THE *Journal* OF
Experimental
Education

Volume 39, Number 4   *Summer 1971*

CONTENTS

Arvil S. Barr, Founder

EDITOR AND PUBLISHER ● 1932-1962

THE PRE-COLLEGE STUDENT SCIENCE TRAINING

PROGRAM OF THE NATIONAL SCIENCE

FOUNDATION: AN EMPIRICAL ANALYSIS [1]

ALEXANDER W. ASTIN
American Council on Education

ABSTRACT

The effects of a summer science program for high school students was evaluated using matched longitudinal data from two national samples of students. Among white students of both sexes, participation in the program appears to increase interest in majoring in science in college, in pursuing a career as a scientist, and in going on for the PhD. degree. Although similar, but less pronounced, effects were observed among black male participants, no such effects were observed among black female participants.

FOR SEVERAL years now the National Science Foundation (NSF) has sponsored a student science training program (SSTP) for high ability secondary school students. The primary purpose of this program is to stimulate the scholary development and scientific interests of science-oriented high school students by means of direct experience with college-level instruction and research. The program also seeks to encourage the development of similar programs through other sources of support.

The purpose of this study was to determine the characteristics of students who are selected to participate in the SSTP and to assess the impact of the program on the student's achievement, career plans, and scientific interests.

NATURE OF THE PROGRAM

According to a recent NSF brochure (4), projects can be designed for either of two types of students: (1) those " from secondary schools in which science instruction is, by national standards, satisfactory or better " or (2) those " with limited educational opportunities who have demonstrated high potential for academic achievement, but in whose secondary schools science training is deficient because of inadequate facilities or instruction." Both types of projects, however, are designed for high-ability students. The emphasis in both programs of study, " should be on substantial and in-

tellectually rewarding programs of study, either through formal class and laboratory work or by means of research programs of suitable difficulty." Although selection of students is the responsibility of the institution submitting the project proposal, the Foundation ordinarily rejects proposals in which the students are drawn largely from a single high school or are enrolled in a regular college summer session course.

A given project ordinarily lasts from 5 to 11 weeks. Although a few students participate during the academic year, the large majority participate during the summer between the junior and senior years in high school. Most projects offer in-depth instruction in specific scientific subjects. A smaller number are more research-oriented, with the student assuming the role of a junior member of a research team under the supervision of a senior scientist. A few projects combine both classroom work and research experience. Instructional costs are borne by the NSF, although most students are expected to provide for their own living and travel expenses. There is, however, financial aid available for students who are unable to participate because of limited financial resources.

EVALUATION OF THE PROGRAM

A crucial consideration in any decision to perpetuate, expand, or terminate educational programs

of this type is their impact on the students who participate. Unfortunately, the opportunities for empirical evaluations of such action programs are typically limited, because the design of the initial program does not provide for a research or evaluation component. In the case of the SSTP, however, an unique opportunity to carry out an empirical evaluation was afforded by the fact that data on large national samples of students were available from two sources: the national testing program of the National Merit Scholarship Corporation (NMSC) and the Cooperative Institutional Research Program of the American Council on Education (ACE). Since the National Merit Scholarship Qualifying Test (NMSQT) is administered during the junior year of high school (before most students participate in the SSTP), and since the ACE's survey is administered during the first few weeks of college (after participation in the SSTP), we can estimate the impact of the NSF Program by examining changes over time in those students who happen to be in both the NMSC and ACE samples.

Selecting the Samples

The first step in selecting the sample for analysis was to identify students who participated both in the survey of entering college freshmen conducted annually by the ACE and in the high school testing program of the NMSC. Although the ACE survey of entering freshmen has been conducted annually since 1966, the class entering college in 1967 was selected since most of these students who had earlier taken the NMSQT would have done so during the spring of 1966, at which time the NMSC also included a brief student questionnaire among its testing materials. This high school questionnaire included questions about the student's vocational interests, educational aspirations, and career plans—information which was also included in the ACE freshman questionnaire of 1967. Such common items of data in the two surveys make it possible to assess changes in the student's interests and career plans between the spring of his junior year in high school and his matriculation to college 18 months later. Since most of these students would have participated in the NSF Program during the summer of 1966 (shortly after the first testing), effects of the program would presumably be revealed by comparing changes in the plans of program participants with those of comparable nonparticipants.

A total of 280,650 entering freshman students at 359 colleges and universities participated in the survey of entering freshmen conducted by the ACE in 1967. In addition to the usual questionnaire information, most of these students also reported the following information which could be used for purposes of identification: name, address, date of birth, state of birth, and sex. These same items of information were also available from the questionnaire completed by students who took the NMSQT 18 months earlier. This latter sample comprised approximately 800,000 students.

In order to identify students who overlapped in the two samples, the two files of identifying information (NMSC and ACE) were sorted in the following order: sex, date of birth, state of birth, last name, first initial, and middle initial. The two sorted files were then matched. Any pair of records that matched exactly in terms of all matching criteria were considered to be from the same student. Although there is a finite probability of a few mismatches under these conditions (students with very common last and first names who were born in very populous states, for example), the number is probably very small. Judging from the small proportion of identical consecutive records in the National Merit File (less than 1 in 5,000), the number of actual mismatches among those assumed to be matches is probably far less than 1 percent. Students who could be matched exactly in terms of most but not all, of the criteria were listed separately for visual inspection. Several additional " matches" were identified in this manner (for example, students who reported a middle initial in one of the testings but not in the other). In all a total of 102,295 matches were identified. Although substantial in size this sample is somewhat below the proportion of freshmen from the ACE sample who would be expected to have taken the National Merit Test (36.4 percent matches as compared to an expected overlap of approximately 55 percent). However, such a loss is to be expected, considering the stringent criteria employed for matching and the relatively high probability that students will not report accurately at least one of the items of identifying data in at least one of the testings. The bias introduced by these stringent matching procedures would, of course tend to exclude students who are reluctant to report complete identifying information as well as students who report such information inaccurately or incompletely.

SSTP participants were identified by means of an item included on the ACE freshman questionnaire A total of 2,018 students (2 percent of the matched sample) indicated that they had been participants in an NSF Summer Program. In addition to these SSTP participants, two " control" groups were selected from the sample of matched Ss: all black students (N= 3,003) and every fifteenth nonblack student (N = 6,484)[2]. The latter subsample, rather than all of the remaining nonparticipants, was selected in order to reduce computing costs.

Characteristics of the Samples

Table 1 shows the sexual and racial composition of the SSTP participants and nonparticipants among the matched Ss and also among all entering college freshmen of 1967 (all data shown are as reported on the ACE freshman questionnaire). The sample of NSF participants is clearly biased in favor of male students. Although such a bias is perhaps to be expected, considering that men typically have stronger interests in science than do women, our data (see Table 1) indicate that the sex ratio among the program participants could be equalized with no appreciable loss in either the level of talent or the degree of science interest of the participating students.

The data in Table 1 reveal an interesting and perhaps unexpected finding concerning the racial composition of the NSF participant group: the proportion of blacks is more than twice as great as it is among the nonparticipants, and also substantially higher than it is among all entering college

TABLE 1

SEX AND RACE OF NSF SUMMER PROGRAM PARTICIPANTS AND NONPARTICIPANTS

	Entering Freshmen Who Took the NMSQT						All Freshmen Entering College In Fall 1967*		
	NSF Participants (N = 2,018)			Nonparticipants (N = 100,277)					
	Male	Female	Total	Male	Female	Total	Male	Female	Total
Percent									
Male			69.7			53.6			55.6
White	91.6	81.5	88.4	94.5	92.2	93.4	90.1	89.6	89.9
Black	5.1	12.7	7.5	2.2	3.9	3.0	3.9	4.8	4.3
Oriental	1.7	3.8	2.3	.9	.8	.9	.9	.7	.8
American Indian	.4	.3	.4	.2	.3	.2	.6	.7	.7

* From Panos, Astin, and Creager (5).
Note: Percentages in each column sum to less than 100 across the four racial categories because a fifth category (" other ") is not shown here.

freshmen. There also appears to be a slight over-respresentation of Oriental students among partici-pants. Considering the fact that blacks typically score below whites on the types of selection criteria normally used in such programs (grades and test scores, for example) (2), the high proportion of black students in the participant group suggests that there may have been a conscious effort to recruit such students into the program. Such recruiting would, or course, be consistent with the second SSTP selection criterion, i.e., to make the program available to students whose opportunities for science training have been limited.

Table 1 also affords an opportunity for us to com-pare the large sample of matched Ss with all fresh-men entering college in the fall of 1967. While the sex ratio among the matched sample is very close to that of all entering freshmen (53.6 percent ver-sus 55.6 percent), the proportion of whites appears to be slightly higher and the proportion of blacks slightly lower than is found among all entering fresh-men. This difference in racial composition could reflect a bias in the schools and students who parti-cipate in the National Merit Scholarship Program, or it might also reflect racial differences in the ac-curacy of reporting the necessary identifying infor-mation.

In view of the sexual and racial biases in the sample of NSF Program participants, all of the anal-yses to be reported subsequently have been perform-ed separately by sex and race. Analyses involving nonblack nonparticipants, however, shall be based on the subsample of 6,484 students rather than on the entire sample of nonblack nonparticipants. Where-ever the analyses involve data from the ACE fresh-man questionnaire, we shall also report national nor-mative data for all entering freshmen of 1967 (5).

Table 2 shows selected background character-istics of the four participant and four nonparticipant groups, as well as national normative data for all entering freshmen of 1967. With respect to age, all groups of NSF Program participants—boys and girls, both black and white—are younger than non-participants. These age differences are especially pronounced among the black males. Thus, although blacks tend to be older than whites among the non-participating boys, the reverse is true among the participants: the blacks are somewhat younger than the whites. A similar trend is observable among the girls, where the proportion of students below age 18 is considerably higher among the black than among the white participants. However, it should be noted that the black girls are more variable with respect to age than are the white girls: proportion-ately more blacks are also above age 18 among the girl participants.

The data in Table 2 also indicate that, without exception, participants come from more highly ed-ucated and more affluent families than do nonparti-cipants. The differences in parental education are somewhat larger in size than the differences in fam-ily income. As might be expected, the education and income levels of the parents of black students are consistently below those of the parents of white students. The differences in family income are par-ticularly striking, with the proportions of families with incomes below $6,000 three to four times greater among the black than among the white students.

An interesting anomoly in the data concerns the relative levels of education of the mothers and fa-thers of the students. Fathers of the white students are consistently more likely than are their mothers to have a college degree. Among the black students,

TABLE 2

BACKGROUND CHARACTERISTICS OF NSF SUMMER PROGRAM PARTICIPANTS AND NONPARTICIPANTS BY RACE AND SEX

Background Characteristic	Boys				Girls				All Freshmen Entering College in Fall 1967*
	White		Black		White		Black		
	Participants	Nonparticipants	Participants	Nonparticipants	Participants	Nonparticipants	Participants	Nonparticipants	
Percent 17 or younger	13.8	6.9	22.7	10.8	13.4	8.8	28.6	15.6	4.8
Percent 19 or older	7.5	11.8	4.5	17.4	4.7	6.6	5.2	9.4	18.5
Parent has a College Degree:									
Percent mothers	36.2	21.6	35.8	18.9	33.5	24.2	31.6	19.2	17.8
Percent fathers	49.6	34.0	30.3	18.0	47.6	37.8	25.0	17.2	26.4
Family Income:									
Percent Below $6,000	9.0	10.9	37.9	42.5	9.8	12.6	39.7	44.3	13.9
Percent $10,000 or higher	65.8	55.8	34.4	25.6	61.1	56.1	29.4	23.0	59.2
Religion of Parents:									
Percent Protestant	52.4	51.7	75.4	62.6	55.8	53.3	73.7	66.7	54.3
Roman Catholic	19.7	31.2	4.6	14.4	24.2	32.1	11.8	12.6	31.7
Jewish	20.2	9.8	.0	.3	12.5	9.1	.0	.0	5.4
"none"	4.3	2.7	.0	1.7	4.5	2.1	.0	1.1	2.1
Percent Living in:									
New England	4.8	7.5	.0	3.0	6.8	8.3	.0	1.8	5.7
Middle Atlantic	30.0	25.1	16.4	26.9	25.0	25.4	21.1	27.0	24.4
North Central	30.4	37.7	25.4	27.8	35.3	36.8	28.9	27.9	39.6
Northwest	6.1	5.7	.0	1.4	7.8	7.7	.0	1.0	5.1
South	20.3	17.3	56.7	37.1	14.9	13.6	46.1	38.3	15.2
West and Southwest	8.4	6.6	1.5	3.8	10.0	8.1	3.9	3.9	9.3

* From Panos, Astin, and Creager (5)

TABLE 3

NMSQT SELECTION SCORES OF NSF SUMMER
PROGRAM PARTICIPANTS AND
NONPARTICIPANTS BY SEX AND RACE

Group	Percentage Scoring Above 105		Percentage Scoring Above 130	
	Partici-pants	Nonpar-ticipants	Partici-pants	Nonpartici-pants
All Fresh-men Enter-ing College Fall 1967*		50.0		9.1
Boys				
White	93.0	70.1	65.1	22.1
Black	41.8	25.2	10.3	3.4
Girls				
White	89.9	65.1	57.7	20.4
Black	39.0	19.3	5.2	1.9

*Estimated national norms from Astin (1). Data
for all other "nonparticipant" groups are unweight-
ed tabulations from those entering freshmen of 1967
(5) for whom NMSQT scores could be obtained.

however—participants and nonparticipants alike—the
mothers are slightly more likely than are fathers to
have a college education. A similar trend has been
observed by Bayer and Boruch (2) in a recent study
of black college freshmen.

Table 2 shows some interesting biases in the
religion of the students' parents. Students from Ro-
man Catholic parents are consistently underrepre-
sented in all participant groups. Among the white
students, parents with no religious preference and
(particularly among the men) parents who are Jew-
ish are overrepresented in the participant group,
whereas Protestants are approximately equally re-
presented in the participant and nonparticipant group.
Among the black students, Protestants are some-
what overrepresented among the participants.

The final category of data in Table 2—the geo-
graphic region of the students' home towns— also
shows certain differences between participants and
nonparticipants as well as between the different races.
Participants, for example, appear to be overrepre-
sented in the Middle Atlantic and Southern states,
and somewhat underrepresented in the New England
states. These biases are particularly evident among
the black students, where the participants are much
more likely than the nonparticipants to come from
the Southern states.

A final item from from Table 2 should be noted.
In comparison with all freshmen entering college in
1967, our sample of matched Ss appears to be some-
what younger and to have somewhat more highly

educated parents than the typical college freshman.
Although the distributions of family incomes and
home states are quite similar in the matched sam-
ple and in the national sample of college freshmen,
the proportion of Jewish parents among the matched
sample is substantially higher than the proportion
in the entering-college population. It seems likely
that these and other differences in race, age, and
parental educational level, are in part the result of
biases in the secondary schools and students who
participate in the National Merit Program.

Selection scores of the various groups on the
NMSQT are shown in Table 3. Rather than simply
computing means for each of the groups, we have
selected two cutting points—105 and 130—to show
the percentage of students in each group scoring
above each of these points. The score of 105 is ap-
proximately the median (fiftieth percentile) for all
college freshmen, and a score of 130 is near the cutting
score used by the NMSC to award letters of com-
mendation and certificates of merit to the partici-
pants. The data show clearly that, within each of
the sexual and racial groups, NSF Summer Program
participants score substantially higher than do non-
participants. In fact, when all four groups of parti-
cipants are combined, nearly 60 percent of the par-
ticipants obtain selection scores above 130. These
findings show clearly that the Program is fulfilling
its objectives in terms of selecting students of ex-
ceptional ability.

The data in Table 3 also reveal small differences
favoring all four groups of boys, and substantial dif-
ferences favoring all four groups of white students.
This latter result reflects the well-known racial
differences in performance on tests of academic
achievement. Nevertheless, it should be stressed
that, among black students of both sexes, SSTP par-
ticipants score substantially higher than do nonpar-
ticipants.

Some of the high school achievements of the var-
ious groups of participants and nonparticipants are
shown in Table 4. Once again, we see here the su-
perior academic and scientific accomplishments of
all four groups of participants. Participants, for
example, are two to three times more likely than
are nonparticipants to make A averages in high
school; very few participants, on the other hand,
make less than a B- average. Membership in
scholastic honor societies follows the same pattern,
with approximately twice as many participants as
nonparticipants being elected to such societies. The
percentage of students receiving National Merit
recognition follows the pattern of test scores re-
ported in Table 3. One unexpected finding here is
the relatively high percentages of black students
who report receiving National Merit recognition;
the percentage is considerably higher than would be
expected from their test scores as shown in Table 3.
One possible explanation here is that the NMSC also
administers the National Achievement Scholarship
Program for outstanding Negro students, in which
the NMSQT is not used as a screening device.

Participants and nonparticipants differ markedly
in terms of winning awards in regional or state
science contests. The rate of such awards in the
participant groups is three to five times greater
than it is in corresponding nonparticipant groups.

TABLE 4

HIGH SCHOOL ACHIEVEMENTS OF NSF SUMMER PROGRAM PARTICIPANTS AND NONPARTICIPANTS, BY RACE AND SEX

| High School Achievement | Boys | | | | Girls | | | | |
| | White | | Black | | White | | Black | | All Freshmen Entering College in Fall 1967 * |
	Participants	Nonparticipants	Participants	Nonparticipants	Participants	Nonparticipants	Participants	Nonparticipants	
Percent With Average Grades of:									
A+, A-, or A	69.1	22.7	30.9	9.3	73.8	32.7	38.1	14.6	14.4
Less than B-	1.9	15.4	14.7	32.3	1.2	7.2	10.5	19.9	30.5
Percent elected to Scholastic Honor Society	74.3	38.7	66.2	31.8	79.0	53.7	71.4	45.8	27.1
Percent receiving National Merit Recognition	51.8	16.9	44.1	14.2	47.2	15.4	28.6	15.3	7.7
Percent Winning Award in Regional or State Science Contest	17.3	3.1	23.5	7.4	12.9	2.5	20.8	6.8	2.5

* From Panos, Astin, and Creager (5).

Table 5 shows the educational and career plans of the various groups as reported in the questionnaire completed when they took the NMSQT during the eleventh grade. These data highlight the very strong scientific interests of the participant groups, as well as their high aspirations for graduate study. It is perhaps surprising that, among the white male students, the percentage of nonparticipants planning careers in engineering (20.1) is not appreciably lower than the percentage of participants planning such careers (21.7). Among the other three groups, however, interest in engineering is somewhat stronger among participants than among nonparticipants. It is also worth noting that the percentages of participants planning careers in "other medical fields" (medical technology, nursing, pharmacy, and so forth) are actually lower than the corresponding percentages of nonparticipants planning such careers. The data in Table 5 also indicate that the career interests of male SSTP participants, as compared to those of female participants, are relatively stronger in engineering and physical science, whereas the career interests of the female participants are relatively stronger in biological science, medicine, and "other medical fields."

In summary, our data on NSF participants indicate that they are an exceptionally able group academically. In addition, their interests in science and aspirations for graduate study are considerably stronger than those of other high school students. They also tend to be younger, and their parents are less likely than other students' parents to be Roman Catholic and more likely to be Jewish or to have no formal religion. Although the parents of SSTP students also tend to be more highly educated and more affluent than typical parents, the group as a whole contains a relatively large proportion of black students. This latter finding is probably attributable to the fact that some of the trainee-ships are intended for students whose educational opportunities have been limited.

Evaluation Criteria

Criteria for assessing the impact of participation in the SSTP were derived from the ACE freshman questionnaire. Two considerations were used: (a) relevance of the questionnaire item to the objectives of the program; and (b) availability of appropriate "pretest" data from the NMSC questionnaire. (The

TABLE 5

ELEVENTH GRADE EDUCATIONAL AND CAREER PLANS OF NSF SUMMER PROGRAM PARTICIPANTS
AND NONPARTICIPANTS BY RACE AND SEX

| | Boys | | | | Girls | | | |
| | White | | Black | | White | | Black | |
Plans	Parti-cipants	Nonparti-cipants	Parti-cipants	Nonparti-cipants	Parti-cipants	Nonparti-cipants	Parti-cipants	Nonparti-cipants
Percent Planning Career in:								
Engineering	21.7	20.1	26.5	17.4	2.8	.2	1.3	.8
Physical Science	30.7	9.3	16.2	10.9	22.1	3.7	18.2	6.5
Biological Science	7.8	2.6	8.8	3.7	14.0	2.1	7.8	2.3
Medicine (MD)	10.2	6.3	14.7	6.8	9.2	2.2	11.7	4.0
Other Medical Field	2.2	5.1	1.5	4.7	7.9	12.0	7.8	12.2
Percent Liking Scientific Research "some" or "much"	92.5	66.8	90.1	67.1	91.6	57.3	80.3	59.1
Percent Planning: PhD Degree	55.7	21.5	52.9	30.1	29.6	7.7	32.5	21.5

latter requirement was considered essential to enable us adequately to control for initial pre-program differences between participants and nonparticipants.) Following these guidelines, five evaluative criteria were selected:

1. Intention to major in a scientific field. This criterion was derived from sixty-six college majors listed in the freshman questionnaire. All students who checked either engineering or one of the physical, biological, social, or behavioral sciences as their "probable field of study" were assigned a score of 1; all other students received a score of 0.

2. Intention to pursue a career in science. This dependent variable was derived from forty-four possible careers listed in the freshman questionnaire. A student was considered to be pursuing a career in science (score 1), (a) if he checked "scientific researcher" as his preferred career, or (b) if he was intending to major in science (1, above) and checked either secondary school teacher or college teacher as his "probable future career." All other students (including those majoring in science and not planning a career in either teaching or research) received a score of 0.

3. Intention to obtain the PhD degree (scored dichotomously.)

4. Placed (first, second, or third) in a state or regional science contest (scored dichotomously).

5. Interest in "making a theoretical contribution to science." (Scored on a 4-point (4-1) scale: Essential, Very Important, Somewhat Important, and Of Little or No Importance.)

Pretest ("Control") Variables

Pretest or control variables included the selection score and the five subtests from the NMSQT, in addition to the following eighty-five measures from the student questionnaires: high school grades, age, highest degree sought (four dichotomies,) father's education, mother's education, parental income, parental religion (four dichotomies), region of residence (five dichotomies), first and second choices of probable major fields in college (34 dichotomies), probable career choice (15 dichotomies) and degree of interest in eighteen specific occupations (scores on eighteen 5-point scales). With the exception of age, residence, and parental religion and income, which were taken from the ACE questionnaire, all items were based on the students' responses to the National Merit questionnaire administered when they took the NMSQT.

Statistical Analyses

The principal technique used to assess the impact

of SSTP participation was multiple stepwise linear regression analysis. For these analyses a 5 percent sample of the 102,295 matched Ss (N = 5,114) was selected. Five separate stepwise analyses were performed, one for each evaluation criterion. The eighty-five pretest measures served as independent variables in each stepwise analysis. Each of the analyses was continued until no additional pretest variable was capable of producing a reduction in the residual sum of squares exceeding p = .05.

The object of these regression analyses was to identify all pretest variables which affected the student's scores on each evaluation criterion, regardless of participation in the SSTP. Once the appropriate weights for each of the pretest variables had been identified, they could be used to equate the various groups of participants and nonparticipants statistically in terms of their characteristics as high school juniors.

RESULTS AND DISCUSSION

Table 6 summarizes the results of the five step-

TABLE 6

SUMMARY OF STEPWISE REGRESSION ANALYSES (N = 5,114)

Evaluation Criterion	R	Number of Pretest Variables entering Equation*	Pretest Variables Receiving Largest Weights
Intention to major in a scientific field	.567	32	Initial major and career plans, interest scales, NMS-QT math
Intention to pursue a career in science	.630	32	Initial major and career plans, interest scales
Intention to obtain the PhD.	.463	27	Initial degree plans, NMS-QT selection sex (male)
Placed in state or regional science contest	.583	20	High school grades, NM-SQT selection, sex (male)
Interest in "making a theoretical contribution to science"	.571	28	Interest in research, initial major and career choice, sex (male), initial degree plans

* Cutoff point was p = .05 (F < 4.00).

TABLE 7

EFFECTS OF PARTICIPATION IN NSF STUDENT SCIENCE TRAINING PROGRAM ON PLANS TO MAJOR IN A SCIENTIFIC FIELD IN COLLEGE

Student Group	N	Percent of Students Planning to Major in a Science Field		
		Expected Based on Spring 1966 data	Actual From Fall 1967 data	Difference Actual - Expected
Participants				
White Boys	1,339	63.7	70.6	+6.9
White Girls	534	49.9	61.6	+11.7
Black Boys	68	67.5	64.3	-3.2
Black Girls	77	50.5	45.5	-5.0
Nonparticipants				
White Boys	3,501	41.0	41.2	+0.2
White Girls	2,983	21.8	22.9	+1.1
Black Boys	1,179	50.5	40.8	-9.7
Black Girls	1,824	36.8	32.6	-4.2

wise regression analyses. The multiple correlation coefficients indicate that the five evaluation criteria can be predicted over the eighteen month interval with moderate accuracy. In general, pretest variables that received the largest weights were those whose content most resembled the content of the evaluation criteria being predicted. Thus, initial major field plans received the largest weights in predicting freshman choice of a major in college. Similarly, degree plans as expressed in the eleventh grade received the largest weights in predicting degree plans at the time of entrance to college.

These regression equations were used to estimate the effects of SSTP participation in the following manner. Weights derived from the stepwise analyses were applied separately to the pretest data of each of the eight groups of SSTP participants and nonparticipants in order to compute the "expected" performance on each of the evaluation criteria. These expected performances (for example, the percentage aspiring to the PhD upon entering college) represent the prediction that one would make from the eleventh grade data, assuming no effect of program participation. These expected scores are then compared with the group's actual performance to determine if the group deviated from expectation. (In the jargon of regression analysis, these difference scores are actually mean residuals from regression.)

Table 7 shows the expected, actual, and difference percentages for the first evaluation criterion (intention to major in a scientific field in college), separately for each of the eight groups. The data indicate that the percentages of white SSTP participants who plan to major in science when they enter college exceeds expectation for both sexes. These differences between the expected and actual percents, when compared with the small differences obtained in both groups of white nonparticipants, indicate that SSTP participation increases the white student's chances of selecting a science major when he enters college.

Results with black students, however, are not so clear-cut. All four groups of black students—participants and nonparticipants alike—show less interest in majoring in science when they enter college than would have been expected from their eleventh grade data. Apparently, there is an interaction between race and changes in science interest over the 18-month interval; that is, the science interests of black students show a somewhat greater decline than those of white students over the 18-month interval. With respect to the effects of SSTP participation, there is some indication that it retards this decline in science interests among black males (-3.2 percent decline) among participants versus -9.7 per-

TABLE 8

EFFECTS OF PARTICIPATION IN NSF STUDENT
SCIENCE TRAINING PROGRAM ON PLANS TO
PURSUE A CAREER IN SCIENCE

| Student Group | N | Percent of Students Planning Careers as Scientists | | |
		Expected Based on Spring 1966 data	Actual From Fall 1967 data	Difference Actual-Expected
Partici-pants				
White Boys	1,339	41.0	49.7	+8.7
White Girls	534	20.3	27.2	+6.9
Black Boys	68	40.1	41.1	+1.0
Black Girls	77	16.6	13.0	-3.6
Nonpartici-pants				
White Boys	3,501	26.5	25.7	-0.8
White Girls	2,983	2.6	2.7	+0.1
Black Boys	1,179	27.7	19.6	-8.1
Black Girls	1,824	6.6	3.5	-3.1

TABLE 9

EFFECTS OF PARTICIPATION IN NSF STUDENT
SCIENCE TRAINING PROGRAM ON PLANS TO
OBTAIN THE PHD DEGREE

| Student Group | N | Percent of Students Planning PHD Degrees | | |
		Expected Based on Spring 1966 data	Actual From Fall 1967 data	Difference Actual-Expected
Partici-pants				
White Boys	1,339	41.8	55.5	+13.7
White Girls	534	24.8	29.5	+ 4.7
Black Boys	68	46.7	52.9	+ 6.2
Black Girls	77	28.6	32.5	+ 3.9
Nonpartici-pants				
White Boys	3,501	21.5	21.5	.0
White Girls	2,983	7.8	7.8	.0
Black Boys	1,179	31.7	30.0	-1.7
Black Girls	1,824	19.9	21.5	+1.6

cent decline among black male nonparticipants) , but there is no indication that it has any such impact among the black women.

The effects of SSTP participation on plans to pursue a career in science are shown in Table 8. Here the pattern of effects is very similar to what we saw in Table 7. Among white students of both sexes, interest in pursuing a career in science appears to be enhanced by SSTP participation. White nonparticipants show the expected difference scores of near zero, but black nonparticipants show a greater-than-expected decline in science interests during the 18-month interval. SSTP participation appears to impede this decline among the black male students, but not among the black females.

Table 9 shows the effects of SSTP participation on plans to obtain a PhD degree. The greatest difference between expected and actual percentages occurs among the white male participants, where the actual score is nearly 14 percentage points above the expected score. This finding indicates that SSTP participation has a pronounced positive effect on the plans of the white male high school student to go on to the PhD degree.

Effects of SSTP participation on PhD aspirations

among the other three groups of participants are not as marked, although the trends are in the positive direction. Once again, there appears to be very little effect of participation among the black female students.

Table 10 shows the effects of SSTP participation on the student's expressed interest in " making a theoretical contribution to science." The difference scores of SSTP participants and nonparticipants indicates that the Program has a positive effect among white students, but only a borderline effect (not statistically significant) among black students. These findings, however, should probably be interpreted with more caution than the findings with the three previous criteria (Tables 7-9), since the National Merit questionnaire did not actually include a " pretest" on this particular item. Thus, the possibility that we have not adequately controlled relevant pretest differences between participants and nonparticipants is greater on this outcome than on the three previous ones.

It should be pointed out that the actual mean scores of SSTP participants as shown in Table 10

TABLE 10

EFFECTS OF PARTICIPATION IN NSF STUDENT SCIENCE TRAINING PROGRAM ON STUDENT INTEREST IN " MAKING A THEORETICAL CONTRIBUTION TO SCIENCE"

Student Group	N	Mean Score on Item *		
		Expected Based on Spring 1966 data	Actual From Fall 1967 data	Difference Actual-Expected
Participants				
White Boys	1,339	2.31	2.50	+0.29
White Girls	534	1.91	2.15	+0.24
Black Boys	68	2.42	2.51	-0.09
Black Girls	77	1.91	1.95	+0.04
Nonparticipants				
White Boys	3,501	1.83	1.80	-0.03
White Girls	2,983	1.34	1.35	+0.01
Black Boys	1,179	2.03	1.90	-0.13
Black Girls	1,824	1.60	1.54	-0.06

* Scored on a 4-point scale: Essential (4); Very Important (3); Somewhat Important (2); Of Little or No Importance (1).

TABLE 11

EFFECTS OF PARTICIPATION IN NSF STUDENT SCIENCE TRAINING PROGRAM ON WINNING AN AWARD IN A REGIONAL OR STATE SCIENCE CONTEST

Student Group	N	Percent Winning an Award in Regional or State Science Contest		
		Expected Based on Spring 1966 data	Actual From Fall 1967 data	Difference Actual-Expected
Participants				
White Boys	1,339	6.9	17.3	+10.4
White Girls	534	6.3	12.9	+ 6.6
Black Boys	68	10.5	23.5	+13.0
Black Girls	77	9.2	20.8	+11.6
Nonparticipants				
White Boys	3,501	3.6	3.1	-0.5
White Girls	2,983	2.6	2.5	-0.1
Black Boys	1,179	7.4	7.4	0.0
Black Girls	1,824	6.8	6.9	+0.1

fall between 2 (" of some importance") and 3 ("very important"). Considering the high proportion of science majors among the SSTP participants (about two-thirds; see Table 7), one might have expected higher scores on this scale. However, in view of the fact that less than half of these SSTP students actually planned a career in science (see Table 8), it is perhaps to be expected that many of them will not attach much importance to theoretical work.

The effects of SSTP participation on the fifth evaluation criterion—winning an award in a regional or state science contest—are shown in Table 11. This is the only one of the five criteria where the Program shows a positive effect for all four groups of participants. Thus, the differences between expected and actual percentages for program participants are all in the positive direction, whereas the differences for each of the four control groups of nonparticipants are very near zero. One note of caution, however, should be added in evaluating these findings: there is some possibility that SSTP participation is for some students the result, rather than a cause of having received an award in a science contest. Some students, for example, may have had their science projects well underway by the time they were considered for participation in SSTP. Such projects would, in turn, make these students more visible to the persons

responsible for selecting SSTP participants.

ALTERNATIVE ANALYSIS

The evidence that SSTP participation increases the student's interests in majoring in a scientific field (Table 7) raises a further question: Does the Program operate to affect interest in all science fields, or only in certain ones ? In order to explore this question, a somewhat different method of analysis was used. Instead of performing regression analyses separately for each of the specific science fields, we decided instead to perform a matching study. While matching as a quasi-experimental technique is generally inferior to regression analysis (3), we chose matching because it permitted us to obtain for each SSTP participant a matched nonparticipant whose eleventh-grade choices of fields of study and careers were exactly the same.

The procedures for obtaining these matches were as follows. The data files for SSTP participants (n = 2,018) and nonparticipants (n = 100, 277) were sorted separately in the following order: sex, initial career choice, initial major field choice, initial degree plans, high school grades, NMSQT scores, and interest scale scores. For

TABLE 12

PERCENTAGES OF STUDENTS CHOOSING VARIOUS FIELDS OF STUDY BEFORE AND AFTER PARTICIPATING IN NSF SUMMER PROGRAM

Major Field Choice	NSF Participants		Matched Controls *	
	Eleventh Grade	Enter College	Eleventh Grade	Enter College
Physical Sciences	34.2	33.1	33.5	26.7
Engineering	16.3	17.2	16.6	16.6
Biological Sciences	13.0	12.8	12.1	9.7
Premedical	9.7	9.4	9.9	10.2
Other Medical	3.8	3.8	3.5	3.4
Social Science	1.9	5.0	1.7	5.2
Undecided	7.5	1.5	7.8	1.8
Other non-science	13.6	17.2	14.9	26.4

*Controls have been matched one-to-one on the basis of sex, initial career choice, initial major field choice, initial degree plans, high school grades, aptitute test scores, and initial interest scale scores.

each SSTP participant, a matched control S was selected from the file of 100,277 participants. In those few cases where an exact match could not be found, the closest match was used. Matching criteria were relaxed in the reverse of the order shown above (that is, initial interest scale scores were relaxed first).

Specific major field choices of the SSTP participants and matched controls are shown in Table 12, both at the eleventh-grade level and also at the time they entered college. The columns of data under " eleventh-grade" indicate that the matching was very close. With the exception of " other " fields (where the difference between participants and controls was 1.3 percent), no specific major field choice showed a difference as great as one percent. (Of course, it probably would have been possible to obtain exact matches in these distributions of eleventh grade choices if initial major field choice had been the first, rather than the third, matching criterion; see above).

The data in Table 12 indicate that the effects of SSTP participation on choice of a major field are limited to a major in either physical or biological science. The decline in student interest in these fields among the matched controls (-9.2 percent) is much greater than it is among the SSTP participants (-1.3 percent). This comparative net gain among the SSTP participants (approximately 8 percent) is very close to the net gains shown previously among white participants in the regression analysis (Table 7). It is important to note that SSTP participation appears to have no effect on student interest in majoring in engineering, medical sciences, or social sciences. The increase in student interest in social science among the SSTP participants, for example, is paralleled by an almost identical increase among the matched controls. The relatively large increase in " other" choices among the matched controls appears to be primarily the result of dropouts from the physical and biological sciences.

A final note of interest from this matching analysis is that the proportion of black students among the matched controls (1.8 percent) was even smaller than the percent among nonparticipants shown earlier in Table 1 (3.0 percent). Thus, when SSTP participants are compared with nonparticipants with identical interests and abilities, the overrepresentation of blacks among the participant group (7.5 percent) appears even larger.

SUMMARY

The purpose of this report was to evaluate the SSTP of the NSF by examining the characteristics of students selected into the program and by estimating some of the effects of the Program on the student's educational and vocational plans and achievements. Longitudinal data from a national sample of students who participated in both the 1966 National Merit Testing Program and the ACE survey of college freshmen in 1967 reveal the following:

1. In terms of academic ability and academic achievement in high school, SSTP participants

represent a distinctly superior group. Compared with other high school students, SSTP participants are also younger, more highly motivated for graduate study, and more likely to be men. Their parents, compared with the parents of other high school students, tend to be more highly educated and affluent, are less likely to be Roman Catholic, and are more likely either to be Jewish or to have no formal religion.

2. The number of black students among SSTP participants is four times greater than the number that would be expected in any comparable group with similar interests and abilities. It seems likely that this overrepresentation of blacks is a direct consequence of the fact that the program attempts to select a portion of the students because their educational opportunities have been limited.

3. Among white students of both sexes, SSTP participation appears to have a positive effect on their interest in majoring in science in college, on their interest in pursuing a career as a scientist, and on their intention to get a PhD. Similar, but less pronounced, program effects were observed among black male participants. Among the black female participants, however, no such effects were observed.

4. The study also produced evidence that SSTP participation increases the student's chances of winning an award in a state or regional science contest while in high school. This finding obtained for black and white students of both sexes.

FOOTNOTES

1. Not for quotation without permission of the author. This study was supported in

part by grant GR-57 from the National Science Foundation.

2. Nonblack students are mostly (97 percent) Caucasian, although they also include small percentages of Orientals and American Indians. For simplicity, however, we shall refer to this group as "white," rather than use "nonblack."

REFERENCES

1. Astin, Alexander W., Predicting Academic Performance in College, The Free Press, New York 1971.

2. Bayer, Alan E.; Boruch, Robert F., " Black and White Freshmen Entering Four-Year Colleges," Educational Record, 371-386, Fall 1969.

3. Campbell, Donald T.; Stanley, Julian C., Experimental and Quasiexperimental Designs for Research, Rand McNally, Chicago, 1966.

4. National Science Foundation, Student Science Training Program (Pre-College) for High-Ability Secondary School Students, National Science Foundation (Document E 69-p-21), Washington, D. C., 1969.

5. Panos, Robert J.; Astin, Alexander W.; Creager, John A., " National Norms for Entering College Freshmen—Fall 1967, " ACE Research Reports, 2: (No. 7), 1967.

BOOK REVIEWS

Robert E. Clasen

book review editor

READING TESTS AND REVIEWS

Buros, Oscar Krisen, Editor, (Highland Park, New Jersey: The Gryphon Press, 1968) , XXII+520 pp. $15.00.

THE COMPLETE title of the volume is Reading Tests and Reviews Including a Classified Index to the Mental Measurements Yearbooks. To the reviewer it is one of the best organized and indexed texts in the field of education. The book is a monumental work by which it is possible not only to obtain information about reading tests of all levels, but also to find sources of data evaluating measures in practically every area of the instructional realm.

Primarily, the Buros book reproduces the reviews on reading tests as originally published in his six Mental Measurement Yearbooks appearing from 1938 to 1965. More than 150 educational experts contributed one to a half-dozen or more extensive reports on the characteristics of tests constructed in the English-speaking world. In addition to tests originating in the United States, there are numerous measures from Canada, England, Australia, New Zealand, and Africa. The reviews are comprehensive as they follow an organizational outline designed by Buros. These nationally-known reviewers are listed at the beginning of the book.

The Preface describes the history of the book extending back to 1940 when plans for the book were first announced. Here will be found a statement of the full scope of the volume with all data brought up to May 1, 1968.

The Introduction explains how Reading Tests and Reviews provides test users with critical information about the merits and limitations of measures that have been classified with respect to reading as "General," "Diagnostic," "Miscellaneous," "Oral," "Readiness," "Special Field," "Speed," and "Study Skills." A Reading Test Index, utilizing the above headings, prints the names of all reading tests reviewed in one or more

AN ANALYSIS OF A SPANISH TRANSLATION

OF THE SIXTEEN PERSONALITY

FACTORS TEST

PATRICK BERTOU
Instituto Venezolano de Investigaciones Cientificas, Caracas, Venezuela

ROBERT E. CLASEN
The University of Wisconsin

ABSTRACT

A Spanish translation of the Sixteen Personality Factors[1] test was administered to 524 freshmen at Venezuela's Central University. The responses were scored using the original scoring key provided by the editor of the test: (1) The results of the test were submitted to an item analysis program (FORTAP). The results of the item analysis showed that only four factors (C, H, F, Q_4) had acceptable reliability (Internal Consistency = +.50). (2) The acceptable factors (C, H, F, Q_4) and items within these factors susceptible to improvement were analyzed from two points of views, verbal content and statistical indices (biserial correlations, X50 and Beta). (3) A Reciprocal Averages Program was used to explore the possibility of improving the factor reliability through changing the weighted values of the response alternatives.

These results indicate that the correct grammatical translation from English to Spanish of the Sixteen Factor Personality Test was not sufficient to obtain acceptable reliability indexes. Item analyses were useful in detecting faulty items.

ONE OF THE crucial problems confronting education in most developing countries is the scarcity of appropriate instruments for locating students on the various relevant subject-matter, personality, attitudinal, aptitude, or interest continua. Because many scientists in developed countries believe that there is essentially no point in repeating instrument development already completed in their countries, there is a great temptation to recommend the translation of desired instruments and to utilize them with the scoring procedures and norms of the country of their development. Secondly, many scientists in developing countries who feel they need tests of one form or another, but who have neither the time nor the resources to develop such an instrument, are tempted to use translated tests and norms as a solution to their problems.

Sufficient data exist (2, 5, 3, 11) to suggest that though national groups are undoubtedly similar in their fundamental humanity, they are also sufficiently different to prohibt the use of identical indices for group and individual comparisons. Nevertheless direct translations of tests are being used without adequate attention to the fact that cultural nuances might require a complete re-norming—perhaps even a complete re-factoring of any given test before it should be used in any of the various educative and mental-health guidance processes.

This study tested the feasibility of using a Spanish translation of Cattell's Sixteen Personality Factor Test (16 PF). For the purposes of this study, the basic factor structure described by Cattell was assumed to be valid; that is, no attempt was made to recreate the basic factor structure. Rather, the study focused upon the function of factors within the battery and upon the function of items within these factors.

Reliability coefficients were considered to be the most important statistical index in determing the acceptability of the factors since reliability coefficients are pertinent to validity in the negative sense, that is unreliable factors cannot be valid (12). Hence, assuming Cattell's 16 PF to be universally valid, if they are not reliable for the Venezulan population tested, it could be stated that the items

and factors do not measure the proposed traits in that context.

For personality tests, the literature reviewed does not establish specific, acceptable values for reliability coefficients. In consequence, an a priori value of .50 was taken as the smallest acceptable value. The decision to use this value was determined on two bases: 1. in the personality test area the reliabilities obtained are frequently lower than those obtained with aptitude and achievement tests (8), and 2. that the lowest reliability value accepted by Cattell in the 16 PF was .54 for factor Q_1 (see Table 1).

PROCEDURES

A translation of the Form A, of the 1962 edition of the 16 PF translated in Chile by Naranjo (10) was adapted to Venezuelan Spanish. Changes were made in the colloquialism of expressions which vary in Spanish speaking countries.

Subjects

The test was given to a sample of 524 Ss composed of three hundred males and 224 females between the ages of 18 and 25. These students were entering freshmen at the Medical School of Venezuela's Central University in Caracas. The test was presented to the students as an experimental questionnaire whose only purpose was to help them in the future if they needed some type of guidance.

Responses were corrected with the two scoring keys provided by the American editors of the test. In all cases the presence of trait answer was scored 2 points, the intermediate response 1, and the absence of trait 0.

Factor reliability, item characteristics (biserial correlations, percent of answer, $X50^2$, and Beta[3]) and possible weighting improvement (reciprocal averaging) were obtained by submitting the data deck to a Fortran Test Analysis Package (Fortap), a University of Wisconsin program developed by Baker and Martin (1). The program was set up to run sixteen times at once, considering each of the factors as an independent test. This was done by instructing the program to give independent results for each set of items included in a factor.

The extent of mutual influence among the sixteen factors of the Spanish version was computed using the Pearson Product Moment Formula. The computational work was done with a computer program developed by Wolfe (13).

RESULTS

The trait construct was accepted at face value. However, in the present analysis, factors which do not meet the minimum reliability criteria level (.50) were not analyzed.

Table 1 shows a comparison between the factor reliability determined in Venezuela with the Spanish version and that obtained in the United States by Cattell. Note that for the Venezuelan sample only

TABLE 1

COMPARISON OF THE RELIABILITY COEFFICIENT OBTAINED IN VENEZUELA AND THE USA*

Factors	Venezuelan Sample	American Sample
A	.28	.82
B	.20	.75
C	.52	.89
E	.45	.82
F	.62	.72
G	.38	.74
H	.71	.70
I	.35	.61
L	.13	.63
M	.14	.79
N	-.01	.64
O	.37	.74
Q_1	-.06	.54
Q_2	.46	.64
Q_3	.38	.61
Q_4	.56	.79

*The reliability coefficients for the American sample were taken from the Manual Handbook (4). They were obtained with 450 young adult males by the split halves method and corrected with the Spearman Brown formula. The reliability coefficients for the Venezuelan sample were calculated using the Hoyt Analysis of Variance method. This method produces results identical to the Kuder-Richardson Formula 20 which represents an average of all possible split halves to obtain a reliability coefficient (9).

four factors (C, F, H, and Q_4) of the Sixteen (4) have acceptable (.50) reliability indices.

Content and Statistical Analysis of the Acceptable Factor

Since only four factors (C, F, H, Q_4) were identified as acceptable, each was analyzed in further detail. To do this, item characteristics within factors were considered and suggestions were made for improving illustrative items.

Factor C. This factor measures dynamic integration and maturity as apposed to general emotion-

ality. It can be equated to Eysenck's "general neuroticism" (6). Usually the C person is easily annoyed by things and people, is dissatisfied with the world situation, his family, the restriction of life, and his own health.

The item analysis (see Table 2) shows a Hoyt reliability of .523 for the thirteen items compris-

TABLE 2

DESCRIPTIVE STATISTICS FOR ITEMS COMPRISING FACTOR C (HOYT r=.523)

Item Number	Weight	Percentage of Responses	r Biserial	X50	Beta
4	2	41	.37	.61	.40
	1	57	-.32	.55	-.34
	0	2	-.38	-5.52	-.41
5	0	19	-.50	-1.73	-.57
	1	10	-.19	-6.77	-.19
	2	71	.49	-1.10	.57
29	0	53	-.55	.16	-.66
	1	8	-.03	-48.69	-.03
	2	39	.58	.51	.72
30	2	73	.38	-1.63	.40
	1	20	-.25	-3.46	-.25
	0	7	-.40	-3.64	-.43
55	2	80	.39	-2.17	.42
	1	11	-.21	-6.03	-.21
	0	9	-.42	-3.14	-.46
79	0	20	-.59	-1.45	-.73
	1	18	-.19	-4.81	-.20
	2	62	.57	-.58	.68
80	0	8	-.55	-2.61	-.63
	1	53	-.20	.34	-.20
	2	39	.41	.67	.45
104	2	66	.29	-1.47	.30
	1	24	-.12	-5.71	-.12
	0	10	-.41	-3.28	-.44
105	2	34	.47	.87	.54
	1	22	.03	25.40	.03
	0	44	-.46	-.35	-.52
129	0	23	-.58	-1.26	-.72
	1	14	-.25	-4.48	-.25
	2	63	.61	-.56	.78
130	2	58	.40	-.48	.43
	1	20	-.09	-9.43	-.09
	0	22	-.43	-1.74	-.48
154	0	6	-.52	-2.92	-.61
	1	27	-.43	-1.38	-.48
	2	67	.57	-.72	.70
179	2	39	.49	.57	.57
	1	28	-.06	-10.19	-.06
	0	33	-.47	-.92	-.53

ing this scale. The mean discrimination index for the three possible choices are:

Type of Choice	Mean Biserial Correlation
Presence of trait (maturity)	.463
In between	-.176
Absence of trait (emotionally)	-.479

These mean biserial correlation values are within the accepted range for discrimination indices.

Looking to the percentages of answers given to each choice (see Figure 1), a higher tendency can be observed toward the use of "presence of trait" choice (53 per cent) than to the "absence of trait" choices (20 per cent).

Analysis of an Item Susceptible to Improvement

Item 104 "When people are unreasonable, I just: (a) keep quiet, (b) in between, (c) despise them."

The Presence of trait, choice "a" has a low Beta (.30) and a negative X50 (-1.47). This is probably due to translation of the choices. "Keep quiet" was translated literally as "don't speak" (me callo) and "despise" as "I disdain them" (los desprecio). This last option is very "strong" in the Venezuelan culture and might be forcing many people low in this scale to select choice "a."

Factor F. This factor is one of the most important in measuring extroversion and introversion. Individuals with high scores in F are usually more optimistic and have a more happy-go-lucky attitude toward life. Individuals with low F scores tend to be more worried and depressed by common life problems. The Hoyt reliability for factor F (see Table 3) is .616. The distribution of individuals choosing between the three possible answers is well balanced 39 per cent, 25 per cent, and 36 per cent (see Figure 1). The mean discrimination indices are the following:

Type of Choice	Mean Biserial Correlation
Presence of trait (extroverted)	.50
In between	-.43
Absence of trait (introverted)	-.49

Analysis of an Item Susceptible to Improvement

Item 83. "I would hate to be where there wouldn't be a lot of people to talk to: (a) true, (b) uncertain, (c) false."

The part of the item "a lot" was left in the Venezuelan version, with the Chilean expression "harta," a word seldom used in Venezuela with this meaning, but frequently used to indicate "being tired of something." This translation error probably explains the poor statistical Beta value (.31) obtained for this item. Changing the word "harta" for "muchos" (many), should improve this item.

Factor H. H minus individuals tend to leptosomatic characteristics (shy and restrained temperament). H plus persons tend to be those individ-

FIGURE 1

DISTRIBUTION OF THE RESPONSES ACCORDING TO CHOICES

	Percentages		
A	40	16	44
C	56	24	20
E	35	27	38
F	39	25	36
G	55	19	26
H	38	27	35
I	39	18	43
L	49	20	31
M	41	18	41
N	48	19	33
O	32	24	44
Q_1	46	20	34
Q_2	46	26	28
Q_3	43	26	31
Q_4	33	20	48
$\overline{X}$	43	22	35

_2_ Presence of trait
_1_ In between
_0_ Absence of trait
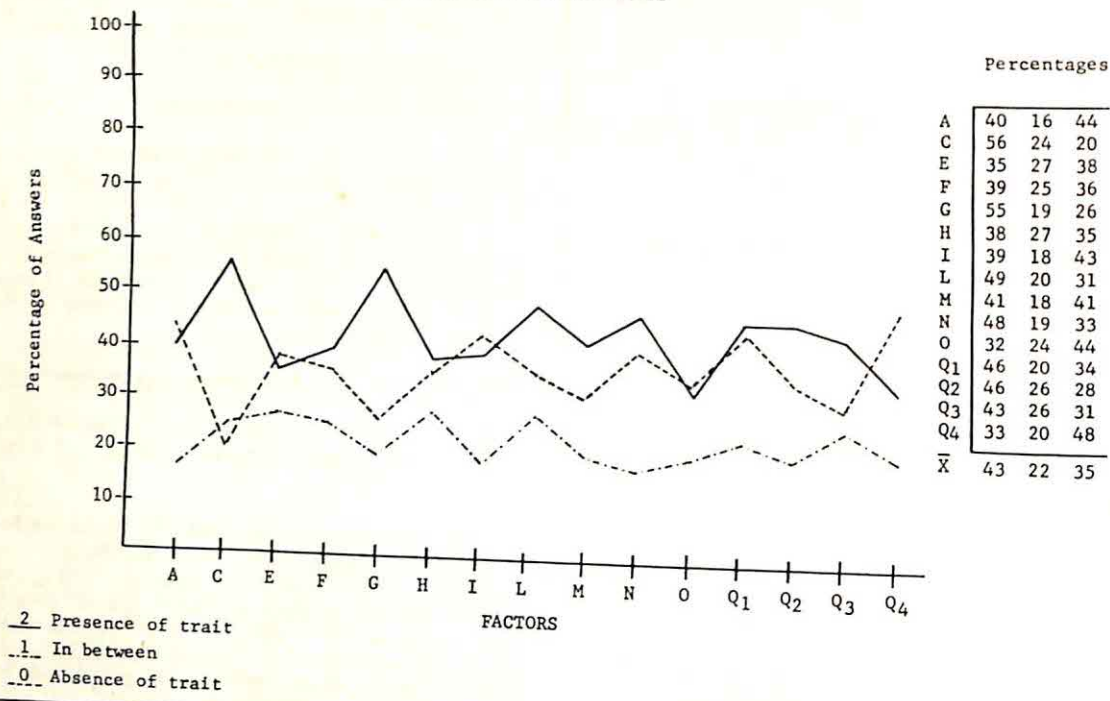
uals who are more adventurous and to like meeting people. This factor is somewhat similar to factor F. The correlation found between these two factors (F and H) is .48.

The Hoyt reliability coefficient obtained for this factor is .709 (see Table 4). This distribution of the answer in percentage value, between the three options, is well-balanced: (a) 38 percent, (b) 27 percent, (c) 35 percent (see Figure 1). The mean biserial correlation for the choices of this factor are:

Type of Choice	Mean Biserial Correlation
Presence of trait (adventurous)	.553
In between	-.034
Absence of trait (shy)	-.541

Analysis of an Item Susceptible to Improvement

Item 86. "I would rather have a job with: (a) a fixed certain salary, (b) in between, (c) a larger salary but dependent on my constantly persuading people I am worth it." Choice "a" intended to indicate shyness, has an X50 of (.77) and a Beta of (-.37). This item undoubtedly has a different connotation in a society not as competitive as the American society.

Factor Q_4. The individuals who score high on factor Q_4 tend to be tense, excitable, and impatient. Those who score low tend to be relaxed and satisfied with their life. The Hoyt reliability coefficient for factor Q_4 is .56 (see Table 5). The distribution of answers tend toward those measuring phleg-

matic behavior (47 percent) (see Figure 1). The mean biserial correlation for the options in this factor are as follows:

Type of Choice	Mean Biserial Correlation
Presence of trait (excitable)	.47
In between	.02
Absence of trait (composed)	-.48

Analysis of an Item Susceptible to Improvement

Item 149. "I tend to tremble or perspire when I think of a difficult task ahead: (a) generally, (b) occasionally, (c) never." This statement is an example of an item that might be good to detect students with emotional problems but it is too "strong" for those that might suffer from anxiety without going to these extremes. From a statistical point of view this item is in the border line of acceptability (X50 = 2.91, B = .45). In the 1968 American edition of the test it was changed to: "I get tense as I think of all the things lying ahead of me."

RESPONSE WEIGHT CHANGES

As stated earlier, another way of improving the reliability of a test is through obtaining a weighting system based not on a priori consideration but on the response records of the individuals.

A reciprocal averages program (RAVE) developed by Baker (1) and available as a library program at The University of Wisconsin was employed. This program does not take into account the score weighted 0. In consequence the original

TABLE 3

DESCRIPTIVE STATISTICS FOR ITEMS COMPRISING FACTOR F (HOYT r=.616)

Item Number	Weight	Percentage of Responses	r Biserial	X50	Beta
8	0	43	-.46	-.39	-.52
	1	18	.61	63.59	.01
	2	39	.46	.59	.52
33	2	42	.52	.41	.61
	1	30	-.04	-11.51	-.04
	0	28	-.56	-1.04	-.68
58	2	26	.40	-.60	.43
	1	62	-.10	2.76	-.10
	0	12	-.43	-2.64	-.48
82	0	26	-.57	-1.14	-.69
	1	21	-.17	-4.74	-.17
	2	53	.58	-.14	.72
83	2	31	.30	1.63	.31
	1	18	.04	21.43	.04
	0	51	-.29	.07	-.31
107	0	37	-.57	-.58	-.69
	1	17	-.06	-15.74	-.06
	2	46	.58	.19	.71
108	0	51	-.33	.09	-.36
	1	24	.10	6.80	.11
	2	25	.32	2.12	.34
132	2	31	.47	1.07	.53
	1	32	.05	9.84	.05
	0	38	-.48	-.65	-.55
133	2	32	.53	.83	.63
	1	23	.03	24.73	.03
	0	45	-.51	-.29	-.60
157	0	70	-.43	1.21	-.46
	1	13	.15	7.42	.16
	2	17	.47	2.04	.53
158	0	38	-.68	-.43	-.94
	1	14	-.07	-14.98	-.07
	2	48	.70	.10	.97
182	2	48	.65	.06	.86
	1	37	-.28	-1.29	-.29
	0	15	-.66	-1.54	-.88
183	2	73	.47	-1.32	.54
	1	17	-.31	-3.08	-.33
	0	10	-.45	-2.89	-.51

scoring system (2, 1, 0) was transformed to 3, 2, 1 to obtain maximum efficiency from the program.

A comparison of the reliabilities obtained with the original weights and with the new weighting system after the RAVE program was applied is given in Table 6. The new reliability indices show a mean difference for the sixteen factors of only .07 (see Table 6) with respect to the reliabilities obtained with the original scoring. An inspection of the new weights shows that most of the changes in scoring were toward giving more weight to the "in between" answers. There was no change in the weighting direction, i.e.; "the presence of trait" choices always received the highest weight, 2 points in the original scoring and 3 points in the new scoring.

TABLE 4

DESCRIPTIVE STATISTICS FOR ITEMS COMPRISING FACTOR H (HOYT r=.709)

Item Number	Weight	Percentage of Responses	r Biserial	X50	Beta
10	2	31	.70	.73	.97
	1	49	-.16	-.03	-.16
	0	20	-.65	-1.31	-.86
35	0	45	-.63	-.21	-.82
	1	16	-.02	-48.04	-.02
	2	39	.66	.40	.89
36	2	59	.57	-.38	.70
	1	21	-.16	-5.10	-.16
	0	20	-.63	-1.32	-.81
60	0	43	-.60	-.30	-.75
	1	27	.09	6.93	.09
	2	30	.60	.90	.74
61	0	52	-.65	.06	-.87
	1	22	.22	3.51	.22
	2	26	.60	1.06	.76
85	0	31	-.68	-.71	-.92
	1	23	-.09	-8.13	-.09
	2	46	.68	.15	.92
86	0	61	-.35	.77	-.37
	1	11	-.03	-47.14	-.03
	2	28	.40	1.37	.44
110	2	57	.48	-.37	.54
	1	20	-.08	-10.81	-.08
	0	23	-.55	-1.36	-.66
111	2	23	.22	3.22	.23
	1	45	.01	23.78	.01
	0	32	-.20	-2.47	-.20
135	2	26	.75	.89	1.13
	1	46	-.06	-1.45	-.06
	0	28	-.64	-.91	-.84
136	2	36	.52	.70	.61
	1	28	-.00	-202.90	.01
	0	36	-.52	-.69	-.60
161	0	52	-.51	.06	-.59
	1	18	.10	9.45	.10
	2	30	.50	1.01	.58
186	2	55	.51	-.23	.53
	1	26	-.25	-2.55	-.27
	0	19	-.44	-1.98	-.43

These results demonstrate that it was not feasible to improve the Spanish version of the 16 PF by changing the original scoring system.

Inter-Factor Correlations

The inter-factor Pearson correlations which resulted from the Spanish version are given in Table

TABLE 5

DESCRIPTIVE STATISTICS FOR ITEMS COMPRISING FACTOR Q4 (HOYT r=.561)

Item Number	Weight	Percentage of Responses	Biserial	X50	Beta
25	0	52	-.44	.13	-.50
	1	11	.21	5.86	.22
	2	37	.36	.94	.39
49	2	35	.52	.72	.62
	1	22	-.01	-72.30	-.01
	0	43	-.49	-.38	-.56
50	2	48	.54	.11	.63
	1	19	-.08	-10.79	-.08
	0	33	-.53	-.84	-.62
74	2	43	.51	.32	.60
	1	9	.02	61.79	.02
	0	48	-.52	-.12	-.60
75	0	47	-.31	-.28	-.33
	1	29	.02	33.07	.02
	2	24	.37	1.84	.40
99	2	32	.59	.78	.73
	1	24	.08	9.29	.08
	0	44	-.60	-.27	-.74
100	0	79	-.46	1.75	-.52
	1	10	.30	4.18	.32
	2	11	.42	2.92	.47
124	2	20	.59	1.43	.73
	1	21	.29	2.66	.31
	0	59	-.64	.32	-.84
125	0	53	-.41	.20	-.45
	1	26	.10	6.17	.11
	2	21	.45	1.80	.51
149	2	11	.41	2.91	.45
	1	38	.20	1.55	.21
	0	51	-.39	.05	-.43
150	0	41	-.36	-.62	-.38
	1	19	-.08	-11.15	-.08
	2	40	.41	.63	.46
174	2	45	.55	.25	.65
	1	23	.04	17.83	.04
	0	32	-.64	-.72	-.82
175	0	48	-.39	-.11	-.43
	1	21	.15	5.60	.15
	2	31	.33	1.52	.35

TABLE 6

HOYT RELIABILITY INDICES OBTAINED WITH THE ORIGINAL SCORING SYSTEM AND WITH THE NEW WEIGHTS DETERMINED BY THE RECIPROCAL AVERAGES METHOD

Factors	Reliabilities From Original Scoring	Reliabilities For the New Scoring
A	.28	.29
B	.20	.20
C	.52	.53
E	.45	.46
F	.62	.63
G	.38	.42
H	.71	.72
I	.35	.38
L	.13	.23
M	.14	.18
N	-.01	-.01
O	.37	.41
Q_1	-.06	-.06
Q_2	.46	.46
Q_3	.38	.38
Q_4	.56	.59

7. In Table 8 are shown the inter-factor correlations obtained by Cattell (3). Comparison of the inter-factor correlations obtained in America and in Venezuela shows several differences:

1. In the American correlation matrix (N = 408), 30 percent of the interfactor correlations are significant at the .01 level. In the Venezuelan matrix (N = 524), 72 percent of the correlations are significant at the same level.

2. Twenty-five percent of the Venezuelan correlation indices are in the opposite sign to the one obtained in America.

3. In the American correlation matrix the highest correlation value is .30. In the Venezuelan matrix, twelve of the correlations have absolute values higher than .30.

In general, these results show that the factor relationship within the original version are substantially different from the one obtained with the Spanish version.

TABLE 7

INTER-FACTOR CORRELATION MATRIX OF THE SPANISH VERSION OF THE SIXTEEN PERSONALITY FACTOR TEST

Factor	A	B	C	E	F	G	H	I	L	M	N	O	Q_1	Q_2	Q_3	Q_4
A	1.00															
B	-.09	1.00														
C	.03	.09	1.00													
E	-.01	.10	.13	1.00												
F	.16	-.01	.17	.42	1.00											
G	.13	-.10	.20	-.21	-.07	1.00										
H	.22	-.06	.42	.41	.48	.04	1.00									
I	.19	-.04	-.20	-.19	-.21	.03	-.11	1.00								
L	-.01	-.04	-.16	.10	.06	-.07	-.02	-.02	1.00							
M	-.01	-.01	-.21	-.13	-.22	-.07	-.17	.31	.09	1.00						
N	.03	.06	.05	-.07	-.13	.08	-.05	.06	-.06	.05	1.00					
O	.06	-.09	-.43	-.11	-.04	-.10	-.31	.08	.18	.10	-.10	1.00				
Q_1	-.13	.06	.15	.01	-.02	-.08	.06	.07	-.04	.00	.09	-.20	1.00			
Q_2	-.23	.01	-.06	-.23	-.43	.01	-.32	.08	-.02	.10	.10	-.02	.11	1.00		
Q_3	.05	.01	.28	-.25	-.17	.26	.05	-.01	-.18	-.10	.14	-.28	.06	.03	1.00	
Q_4	.00	.01	-.53	.04	-.07	-.21	-.32	.13	.26	.21	-.05	.45	-.16	.03	-.35	1.00

CONCLUSION

The results reported herein must be evaluated with the fact that they represent the first attempt made in Venezuela to validate the 16 PF.

The reliability indices obtained show that only four factors (C, F, H, Q_4) have an acceptable internal consistency and in consequence might also have acceptable validity.

The distribution of the answers shows a general imbalance in favor of the "presence of trait" choices (43 percent) over the "absence of trait" choices (35 percent).

The discrimination indices indicate that 38 percent of the items forming the acceptable factors made no substantive contribution to the personality trait they are intended to measure. The same is true for 71 percent of the items forming the nonacceptable factors.

A comparison of the inter-factor correlations matrices obtained in America and Venezuela showed substantive differences. This indicated that the factor relationship employed by Cattell to establish the second order factors may not be valid in Venezuela.

It was demonstrated with a reciprocal averages analysis that it is not possible to improve the statistical indices by employing a different weighting system in the scoring of test results.

One problem of this Spanish version was based on the fact that all the items were written taking as a prime consideration the grammatical similarity of the translated words and sentences, instead of trying to culturally adapt the construct behind the original American items. Another problem was the quality of the 1962 edition, according to a letter received from the editors (Institute for Personality and Ability Testing) of the test, the new 1968 American edition is superior to the edition translated into Spanish. Upon review it was found that sixty-nine items (39 percent) were different in content in the 1968 edition from the 1962 edition.

The statistical results obtained in the present study demonstrate that this Spanish version of the 16 PF does not meet the minimum requirements for reliability and validity established by the American

Psychological Association in its "Standards for Educational and Psychological Test and Manuals" (12); therefore it can not be recommended for the diagnosis, prognosis, or evaluation of personality with the group from whom the data were collected. Further research is necessary to ascertain the external generalizability of this inference to Venezuela or Latin America in general.

In adapting tests from different cultures it is reasonable to suggest that attention must be given to the translation of individual items. Given that test scores in education and psychology are based on collectivities of items, it is crucial to learn more about the cognitive characteristics which influence item responses. To accomplish this, studies which systematically vary the characteristics of the original item in relation to varying respondents are recommended. Finally, test translations must be followed by reliability and validity studies before any attempt to norm the instrument can be sensibly undertaken.

FOOTNOTES

1. Institute for Personality and Ability Testing, Champagne, Illinois.

2. X50 is the point in the criterion scale at which the item choice has maximum discrimination, i.e., it is the point on the criterion scale, given in standard deviation units, corresponding to the median of the item characteristic curve of that choice. Subjects with a criterion score equal to X50 have a 50-50 chance of choosing that response.

3. Beta is the reciprocal of the standard deviation of the item characteristic curve and can be thought of as the slope of the item characteristic curve at the X50 point. It gives the discrimination power of the item in values that go from $\pm$ infinity.

4. It is interesting to note that the two factors (N, Q_1) with the lowest reliability (-.01, -.06) are, according to Cattell, acquired social traits.

REFERENCES

1. Baker, F. B.; Martin, T.J., "A Fortran Test Analysis Package, Laboratory of Experimental Design," Wisconsin Research and Development Center for Cognitive Learning, The University of Wisconsin, 1968.

2. Cattell, R., "A Cross-Cultural Comparison of Patterns of Extraversion and Anxiety," British Journal of Psychology, 1, 3-15, 1961.

TABLE 8

INTER-FACTOR CORRELATION MATRIX OF THE ORIGINAL VERSION OF THE SIXTEEN PERSONALITY FACTOR TEST

Factor	A	C	E	F	G	H	I	L	M	N	O	Q_1	Q_2	Q_3	Q_4
A	1.00														
C	.13	1.00													
E	.07	.06	1.00												
F	-.01	.05	.00	1.00											
G	.00	.04	-.02	.05	1.00										
H	-.36	-.12	-.18	-.26	.00	1.00									
I	-.29	-.04	-.06	.01	.00	.14	1.00								
L	-.01	-.01	-.06	-.11	-.04	.26	-.04	1.00							
M	.03	.14	.06	.26	-.04	.05	-.14	-.07	1.00						
N	.08	.01	-.11	-.12	.13	-.05	.26	-.05	-.02	1.00					
O	.02	.19	.00	-.04	.02	.13	-.11	-.08	-.06	.04	1.00				
Q_1	-.15	.05	-.31	-.02	.11	-.09	.02	.01	-.19	.00	.01	1.00			
Q_2	.07	-.01	.18	.22	.02	.03	-.11	-.01	-.11	.03	.08	-.14	1.00		
Q_3	.30	-.19	.12	.22	-.19	-.29	-.02	.11	-.04	-.22	.20	.00	.12	1.00	
Q_4	.04	.09	.02	.05	-.05	.01	.15	-.15	-.01	-.07	-.17	.03	.01	.23	1.00

3. Cattell, R.,"Validation and Intensification of the Sixteen Personality Factor Questionnaire," Journal of Clinical Psychology, 20, 411-418, 1956.

4. Cattell, R.; Eber, H., Handbook for the Sixteen Personality Factor Questionnaire, The Institute for Personality Testing, Illinois, 1957.

5. DeVos, G.,"Symbolic Analysis in the Cross-Cultural Study of Personality," Studying Personality Cross-Culturally, Kaplan, Bert (ed.), Row, Peterson and Company, New York, 599-636, 1961.

6. Eysenck, H., The Structure of Human Personality, Methuen and Company, London, 1953.

7. Kaplow, B., "Personality Study and Culture," Studying Personality Cross-Culturally, Row, Peterson and Company, New York, 301-312, 1961.

8. Lindquist, E.F. (ed.), Educational Measurement, American Council on Education, Washington, 1963.

9. Lord, Frederik N.; Novik, Melvin, Statistical Theory and Mental Test Score, Addison Wesley Publishing Company, Mass., 1968.

10. Naranjo, C., "Cuestionario para 16 Factores de la Personalidad," Escuela de Medicina, Universidad de Chile.

11. Singer, M., "A Survey of Culture and Personality Theory and Research," Studying Personality Cross-Culturally, Kaplan, Bert (ed.), Row, Peterson and Company, New York, 9-92, 1961.

12. Standards for Educational and Psychological Test Manuals, American Psychological Association, Washington, D.C., 1966.

13. Wolfe, Richard, Computer Program Files, Instructional Research Laboratory, University of Wisconsin, 1967.

Book Reviews

of the six Mental Measurements Yearbooks. A coding system is used by which information is given concerning such basic information as the name of the publisher, for what levels the tests were designed, the organization of the test, and where additional information can be found.

The heart of the book is the set of complete reviews from the six Mental Measurements Yearbooks given in chronological sequence. It puts into one volume the information that otherwise would have to be searched out in six books extending back to 1938.

As an additional useful tool, there is a chapter that presents a classified listing of all tests appearing in one or more of the six Mental Measurements Yearbooks. These tests cover the entire field of objective measurement from "Achievement Batteries" to "Specific Vocations."

Lastly, for ready reference, there are a Publishers Directory, an Index of Test Titles, and an Index of Names of Authors, Test Reviews, and Others Mentioned in the Volume.

Reading Tests and Reviews has been skillfully designed to provide an efficient tool for educators who wish to be intelligently informed about available measures. The book is a delight to use and takes its place along side of Webster's Unabridged Dictionary as a basic reference for the professional.

> Alfred S. Lewerenz, Reviewer
> Educational Consultant
> Hollywood, California

REACH, TOUCH AND TEACH

Borton, Terry, (New York: McGraw-Hill, 1970), 213 pp. $4.95.

THERE APPEARS to be the emergence of a new paradigm in education. The past and present failure of the schools to educate a substantial number of the young is bringing forth not only devastating critiques, but also some exciting proposals for the future. One such contribution, Reach, Touch and Teach is outstanding insofar as it not only provides an image of a future paradigm for education but reveals glimpses of the transformations through which schools and individual teachers must go to implement the new paradigm. For those whose goal is process education, Terry Borton has presented a useful guide.

In the new paradigm, the goal of education is the learning of those processes and skills through which an individual may change himself. In part, Borton notes, this goal is similar to the classical educational objective of training habits of mind; according to that model, the student would develop the intellectual skills to discover

THE RELIABILITY AND VALIDITY OF

QUICK TESTS WITH HIGH SCHOOL SENIORS

GEORGE W. BOHRNSTEDT,   PHILIP LAMBERT, and EDGAR F. BORGATTA
University of Minnesota   The University of Wisconsin

ABSTRACT

The Quick Word Test (QWT), Quick Number Test (QNT), and a number of criterion verbal and numerical tests were related with the English and Math grade point average (GPA) scores in this study. The QWT, in general, had lower correlations with English GPA scores than the criterion tests. The correlations between the QNT and the Math GPA was approximately at the same level as the criterion measures.

IN A PRIOR report, Borgatta and Bohrnstedt (2) indicated the utility of a " Quick" test in assessing the performance of college age students. The report suggested that the correlations of the " Quick" test with GPA was essentially of the same magnitude as those with standard college assessment tests. Since both the Quick Word Test (QWT) (3) and the Quick Number Test (QNT) (5) were designed to be used with the range of intelligence found in a normal adult population, they should be well-suited for use with students in the high school range. From a practical point of view, the " Quick" tests might be useful in localities where more conventional testing to be used in advising students may be more difficult to administer. Thus, the use of tests, rather than the misuse of tests, has been under attack and conventional testing programs have been cut out of budgets. Additionally, in some locations tests are administered on a self-selective basis only to those who ask to be tested. This may exclude some persons who choose not to be tested for economic reasons, but motivational and aspirational factors may be very important in the self selection. Obviously, persons never tested cannot be advised, and thus many persons' high potential may be neglected if additional supporting information is not in their files to confirm high school performance, or to lead to examination for potential in cases where high test scores are not accompanied by high grades. Having some tests uniformly administered in senior high school can be justified by the additional information it puts into the student's dossier. The special interest in " Quick" tests of the type reported in this research is that they can make testing feasible in difficult situations, and less difficult circumstances, can add to the base on which advice is offered to students.

In the current research, cooperation was provided by a school system in a southwestern city in which several types of standard tests commonly used for college advisement were available. The " Quick" tests were administered in three schools, providing 1,186 participants for the study. The tests selected as relevant for comparisons with the " Quick" tests were the American College Test (ACT) (1) English and Math scores, the National Merit Scholarships (NMS) (6) English Usage and Math Usage scores, and the College Entrance Examination Board (CEEB) (4) Verbal and Math scores. Different (but overlapping) subgroups of the total had participated in each of these testings, providing subsamples of 712, 373 and 190 respectively. The QWT was administered with a time limit of 15 minutes, although this is normally a power test. The time restriction was not felt to materially alter the performances of students as at this level virtually all complete the form in 15 minutes.

The most desirable comparison that could be made would be in subsequent performance of the participants. However, aside from not having these longitudinal data available, such data necessarily include only the segment of students that go on for higher education. The best predictor of future performance is commonly identified as " current" performance; thus, in lieu of a projection into the future, the criteria used in this study are the cumulative GPA's in English and Math available for the students.

RESULTS

Table 1 presents the intercorrelation between the ACT English, ACT Math, QWT, QNT, English GPA, and Math GPA for a subsample of 712 students.

TABLE 1

INTERCORRELATION BETWEEN ACT ENGLISH,
ACT MATH, QWT, QNT, AND GPA (N = 712)

		1a	2a	3	4	5	6
1a.	ACT English	—	.48	.62	.36	.68	.50
2a.	ACT Math		—	.39	.74	.49	.68
3.	QWT			—	.38	.54	.38
4.	QNT				—	.41	.63
5.	English GPA					—	.65
6.	Math GPA						—

The ACT English correlates with the English GPA with a coefficient of .68, while the QWT correlation coefficient with the English GPA is only .54. While a correlation of .54 between a verbal test and a GPA in English is substantial, the value of .68 for the ACT English score is impressively large in this sample. The correlation coefficients for the ACT Math and QNT with the Math GPA respectively .68 and .63, both relatively high values.

Table 2 presents the correlation coefficients for National Merit Scholarship English Usage and the QWT with the English GPA, which are respectively .62 and .45. The correlation coefficients for the NMS Math Usage and the QNT with Math GPA are respectively .68 and .62. Thus, the "Quick" tests are performing approximately in the same way in comparison to the NMS exams as in comparison to the ACT, but in the subsample of the 373 students for which the NMS was available, the relationships between the verbal scores and the criterion are somewhat lower.

In Table 3 the data are presented for the subsample of 190 students for whom CEEB scores were available. The data indicate the same pattern, with the CEEB verbal and the QWT correlated to the English GPA with coefficients of .64 and .41 respectively. The CEEB Math and QNT correlated with the Math GPA with coefficients of .66 and .62 respectively.

DISCUSSION

The data presented in this analysis are initially disappointing for the QWT. Possibly, a problem arises in the fact that the Ss of these analyses are, self-selectively, the high performing students in the

TABLE 2

INTERCORRELATION BETWEEN NMS ENGLISH
USAGE, NMS MATH USAGE, QWT, QNT, AND
GPA (N = 373)

		1b	2b	3	4	5	6
1b.	NMS English Usage	—	.46	.47	.38	.62	.53
2b.	NMS Math Usage		—	.26	.69	.43	.68
3.	QWT			—	.37	.45	.32
4.	QNT				—	.38	.62
5.	English GPA					—	.69
6.	Math GPA						—

high school, and thus the relationship between the tests of abilities and the criteria are depressed by the reduced individual variability. Still, the question of why the QWT should by noticeably lower than the other verbal tests is not answered by this general restriction. The average score for the QWT is high (74 out of a possible 100 for the CEEB subsample, for example), and this suggests it might be advisable with such groups to use the High Difficulty Form of the QWT.

The "Quick" tests in this application performed at a level that may be considered reasonable but not outstanding. Other reports have been more favorable, and possibly this more modest performance emphasizes a more universalistic point, namely given the objective of providing more information to use as a basis for advising students, attitional efficient assessment instruments are required. The principle of efficiency can be applied in test development to shorten the time required. Possibly as more efficient tests of ability and performance are developed, they can be administered more routinely as "one more piece of information" in a dossier for students. In general, a major problem that confronts the educational testers is that insufficient general testing occurs because accumulation of such information requires a formidable investment of time and resources. The "Quick" tests or other tests developed with equivalent strategies may facilitate a more mundane view to the collection of such information.

TABLE 3

INTERCORRELATION BETWEEN CEEB VERBAL,
CEEB MATH, QWT, QNT, AND GPA (N = 190)

		1c	2c	3	4	5	6
1c.	CEEB Verbal	—	.71	.54	.36	.64	.48
2c.	CEEB Math		—	.35	.68	.50	.66
3.	QWT			—	.29	.41	.28
4.	QNT				—	.37	.62
5.	English GPA					—	.65
6.	Math GPA						—

REFERENCES

1. American College Testing Program, ACT Battery, ACT, Iowa City, Iowa, 1967.

2. Borgatta, Edgar F.; Bohrnstedt, George W.; "The Use of the Quick Number Test in the Prediction of Academic Performances," Educational and Psychological Measurement, 29:921-925, 1969.

3. Borgatta, Edgar F; Corsini, Raymond J., The Quick Word Test, Harcourt, Brace and World, New York, 1967.

4. College Examination Entrance Board, CEEB Battery, Educational Testing Service, Princeton, New Jersey, 1967.

5. Corsini, Raymond J; Borgatta, Edgar F., The Quick Number Test (QNT), The Journal of Experimental Education, 36:7-10, 1968.

6. National Merit Scholarship Corporation, MNS Battery, NMSC, Evanston, Illinois, 1967.

BLACK PUPILS CAN BE TAUGHT TO LISTEN

PERRY R. CHILDERS
The University of Wisconsin-Milwaukee

ABSTRACT

Sixty-four black seventh grade pupils, matched on intelligence and reading ability, served as Ss for an experiment in critical listening. The Ss were randomly assigned to experimental and control groups, thirty-two in each. The Ss were pretested on listening ability. The experimental group received 2 weeks of instruction and work in critical listening. Both groups were posttested using an alternate form of the pretest instrument. Results demonstrated a mean gain of 8.44 points, significant beyond .01 for the experimental group. The control group results evidenced no gain. It was concluded that the type S used in the study could benefit substantially from systematic instruction relative to the experimental criterion.

THERE IS an increasing awareness of the importance of listening ability in relation to pupil achievement. Investigations by Russell (6), Witty (8), and Duker (2) offer evidence on the importance of critical listening in reading and in written communications. Researches indicate critical listening is an identifiable factor, separate from general verbal intelligence, vocabulary, and reading ability (4), and that such an ability is a modifiable skill (1).

Lundsteen (5) has demonstrated that given appropriate instruction and materials, the elementary pupil has considerable capacity for improved critical listening. Winter (7) and Childers (1) found that there is significant improvement in listening ability from fourth through sixth grade, and that listening ability becomes less a function of measured intelligence through the sixth grade. Fawcett (3) corroborates these findings. Young (9) found a relationship between reading comprehension and hearing comprehension and concluded that much reading disability resulted from poor language comprehension.

While the previously cited studies offer strong evidence of what could be done in individual schools to improve class performance and measured achievement, they cannot be generalized to certain ethnic/racial groups. Secondly, lower socioeconomic levels were not adequately represented. The experiment reported herein was conducted to provide data on culturally disadvantaged black pupils in a large, urban school system.

METHOD

Sixty-four black pupils, matched on measured intelligence and reading level, were randomly assigned to an experimental group and a control group. All students were in seventh grade.

The STEP Listening Comprehension Test, Form 3A, was used to pretest. A profile chart for each student in the experimental group was formed showing apparent strengths and weaknesses. The students were then given special instruction and training in critical listening with an emphasis on areas needing improvement. The program centered on content areas designed to develop such skills as comprehension, interpretation, and critical evaluation. The types of listening activities included narration, exposition, direction, simple explanation, judgment, and persuasion. The experimental group received one lesson per day for 2 weeks (50 minutes per lesson). The control group received no training as they followed the usual class routine. At the end of 2 weeks of instruction, both groups were administered the STEP Listening Comprehension Test, Form 3B.

RESULTS

The mean IQ for the experimental group was 88.84, the control group 91.31 (California Test of Mental Maturity, Form L). The mean reading score for the experimental group was 54.34, the control group 55.28 (Iowa Test of Basic Skills, Form 1). The grade level for the two groups combined was 5.5. Equivalence between the experimental and control groups was assured.

The pretest listening comprehension results for

TABLE 1

COMPARISON OF EXPERIMENTAL AND CON-
TROL GROUPS ON LEVELS OF INTELLIGENCE,
READING, AND AUDING SKILL

	Experimental $\overline{X}$	Control $\overline{X}$	difference
IQ	88.84	91.31	2.47
Reading	54.34	55.28	.94
STEP (A)	35.18	38.09	2.91
STEP (B)	43.62	38.18	5.44**
difference	8.44**	.09	

**Pt $<.01$ df = 62

,the experimental group was 35.18, and for the con-
trol group 38.09. This difference was not signifi-
cant. The posttest listening comprehension mean
score for the experimental group was 43.62, and for
the control group 38.18. This difference of 5.44 was
significant beyond the .01 level. The difference of
8.44 points between pre- and posttest results for the
experimental group was significant beyond the .01
level. There was no difference between pre- and
posttest results for the control group.

CONCLUSION

The results presented support the conclusion that
black, elementary school pupils, designated as eco-
nomically and educationally disadvantaged, have the
capacity for significant improvement in critical lis-
tening ability (auding). Such pupils as those used
as Ss in this study are characteristically one or more
grade levels behind in reading. Their measured in-
telligence test scores (IQ's) are also below the lev-
el usually found associated with suburban Ss and test
booklet norms. The extent of reading retardation is
approximately equal to the percentage of retardation
in measured intelligence.

Paper and pencil intelligence test scores (IQ's)
are a function of reading ability (given). Reading
ability is a function of critical listening skills (giv-
en). Critical listening skills can be significantly
improved in black elementary pupils (demonstrat-
ed). Therefore . . . improved reading achievement
and/or improved scores on paper and pencil tests of
intelligence in conjunction with a program of system-
atic instruction in listening skills development ap-
pears as a logical next step in this line of research.

REFERENCES

1. Childers, P. R., "Listening Ability is a Modi-
fiable Skill," The Journal of Experimental Ed-
ucation, 38:1-3, Summer 1970.

2. Duker, Sam, "Listening and Reading," Ele-
mentary School Journal, 65:321-29, March 1965.

3. Fawcett, Annabel E., "Training in Listening,"
Elementary English, May 1966, pp. 473-76.

4. Lundsteen, Sara W., "Teaching and Testing
Critical Listening in the Fifth and Sixth Grades,"
Elementary English, 41:743-47, November 1964.

5. Lundsteen, Sara W., "Critical Listening: An
Experiment," Elementary School Journal, March
1966, pp. 311-15.

6. Russell, David H., "A Conspectus of Recent Re-
Research on Listening Abilities," Elementary
English, 41:262-67, March 1964.

7. Winter, Clotilda, "Listening and Learning,"
Elementary English, October 1966, pp. 569-572.

8. Witty, Paul A., "A 1964 Study of TV Compari-
sons and Comments," Elementary English, 42:
134-41, February 1965.

9. Young, William E., "The Relation of Compre-
hension and Retention to Hearing Comprehension
and Retention," The Journal of Experimental
Education, 5:30-39, 1936.

ANALYSIS OF VARIANCE AND LATIN SQUARE

PROBLEMS BY MULTIPLE REGRESSION ANALYSIS

LEVERNE S. COLLET and JAMES H. MAXEY
The University of Michigan

ABSTRACT

The purpose of this paper is to provide concrete illustrations of the efficacy of the multiple regression approach to the analysis of experimental results. A bridge between theoretical and specific applications is provided by parallel multiple regression and analysis of variance solutions for two typical educational designs. Detailed illustrations are given in the techniques of writing linear constraints and determining degrees of freedom. The major advantages of the multiple regression approach are its adaptability to unusual designs and the facilitation of meaningful interpretation afforded by the provision of regression weights in addition to the usual F ratios.

BOTTENBERG AND WARD (1) and Cohen (2) describe multiple regression (MR) as a very powerful and flexible tool which deserves much wider use among researchers. They point out the MR can be used to analyze any data for which analysis of variance is appropriate. The general equivalence of the two methods is illustrated in Figure 1.

FIGURE 1

PARTITIONING SUMS OF SQUARED DEVIATIONS ABOUT THE GRAND MEAN

	Total	Error Within Groups	Treatment Between Groups
ANOVA	$\Sigma\Sigma(Y_{ij}-\bar{Y}..)^2$	$= \Sigma\Sigma(Y_{ij}-\bar{Y}._j)^2$	$+ n\Sigma(Y._j-\bar{Y})^2$
MR	$\Sigma\Sigma(Y_{ij}-\bar{Y}..)^2$	$=(1-R^2)\Sigma\Sigma(Y_{ij}-\bar{Y}..)^2$	$+R^2\Sigma(Y_{ij}-\bar{Y})^2$

An advantage of MR is that its solution routines are unaffected by unequal N's or incomplete block designs—both of which are beyond the capacity of many ANOVA programs. However, a search of the literature revealed a dearth of detailed MR solutions for practical research problems. It is the intent of the authors to bridge some of the gaps between theory and specific applications by providing comparative

ANOVA and MR solutions to two typical problems. A simple 2-factor design and a 3x3 latin square are discussed. A wide variety of designs can be solved by simple extension of the illustrations given.

BACKGROUND MATERIAL

The key to understanding the MR solution comes in the recognition that membership in various treatment groups can be represented as dichotomized variables. Consider just two groups G_1 and G_2. Membership or nonmembership in these two groups is illustrated in Table 1.

Notice that Ss 1 and 2 belong to group 1 and Ss 3 and 4 belong to group 2. This linear constraint setup for MR is equivalent to saying that Ss 1 and 2 belong to treatment A and Ss 3 and 4 belong to treatment B in an ANOVA problem.

The formula for the F test using the MR solution is:

$$F = \frac{(R^2_{Y.AB} - R^2_{Y.A}) / df_1}{(1-R^2_{Y.AB}) / df_2}$$

This formula tests any increment to $R^2_{Y.A}$, due to the addition of B. The df_1 for the numerator is the number of linearly independent vectors in the

TABLE 1

MEMBERSHIP OR NONMEMBERSHIP

Subject	G_1	G_2
1	1	0
2	1	0
3	0	1
4	0	1

full model minus the number of linearly indepen-
dent vectors in the restricted model. The df_2 for
the denominator is N minus the number of linearly
independent vectors in the full model. Due to the
calculation algorithm the identity vector is always
automatically included and should be counted as one
of the vectors for both the full and restricted mod-
els. It is important to to remember that the d f
are exactly the same as in a traditional analysis
using the conventional F tests.

The three main difficulties in using MR solu-
tions are:

(1) Learning to write the proper linear con-
straints for a given problem so that they re-
present all the different combinations of
group membership and interactions.
(2) Learning to choose the multiple correlations
which test the specified hypothesis, and de-
termining the degrees of freedom associ-
ated with each comparison.
(3) Learning to use a computer program to an-
alyze data. The program used in this study
was Veldman's (3) Program Regran. The
MR or General Linear Hypothesis programs
from the Biomedical series are other possi-
ble choices.

Studying the following sample problems, which il-
lustrate these three difficulties, offers a method
whereby they can be recognized and overcome.

EXAMPLE PROBLEM: 2-WAY ANOVA

A numerical example of a 2x2 factorial exper-
iment having ten observations per cell will be used
to illustrate the computational procedures. Sup-
pose that an experimenter is interested in evalua-
ting how two methods of teaching (factor A) affect
changes in achievement in two categories, boys,
and girls (factor B). The dependent variable is
an achievement test using gain scores. The forty
Ss have been randomly assigned to one of the four
groups.

The analysis was done by using Veldman's (3)
AVAR23 program. The notation used, (Table 2)

TABLE 2

NOTATION TABLE

	Boys B_1	Girls B_2
Method A_1	G_1	G_2
Method A_2	G_3	G_4

TABLE 3

OBSERVED SCORES

G_1 (A_1, B_1)	G_2 (A_1, B_2)	G_3 (A_2, B_1)	G_4 (A_2, B_2)
23	30	42	31
31	60	31	13
42	57	45	18
23	27	52	22
54	38	28	23
72	62	21	41
81	47	17	37
93	43	31	18
72	37	18	17
67	48	12	16

observed scores, (Table 3), summary table of
means, (Table 4) and two-way Analysis of Variance
Source Table (Table 5) follow.

Based on these results one can conclude that
there was a significant difference between teaching
methods at the .0001 level and that there was no
significant difference between sexes and no signifi-
cant interaction if a .05 alpha level is used by the
researcher.

TWO-WAY ANALYSIS USING MR

The same problem will now be analyzed using
the MR technique. In setting up the linear con-
straints there are several things to consider, either
membership or nonmembership in the various fac-
tors and the criteria. A S is assigned a 1 if he is
a member of a certain level and a -1 if he is not a
member of that level. This gives the set of linear
constraints shown in Table 6.

Since in the presence of the identity vector the
scores for A_2 are completely determined by A_1 and
B_2 determined by B_1, it is important to eliminate
this redundancy in order to have a non-singular ma-
trix. Variable X_1 represents factor A and variable
X_2 represents factor B. Variable X_3 represents
the AB interaction term and is generated from the
product of X_1. X_2. Variable X_4 is the criteria
score. If one desires, 0's may be substituted for
the -1's. Table 7 shows how members of each of
the four groups are coded and is the raw data that
goes on the computer cards.

To compute the F ratio for the main effect of A,
variables X_1, X_2, X_3 were used to predict X_4, then

TABLE 4

AB SUMMARY TABLE OF CELL AND MARGINAL
MEANS

	b_1	b_2	
a_1	55. 8	44. 9	50. 35 = A_1
a_2	28. 7	23. 6	26. 65 = A_2
	42. 75	34. 25	
	B_1	B_2	

TABLE 5

TWO WAY ANALYSIS OF VARIANCE SOURCE TABLE

SOURCE	S.S.	D.F.	M.S.	F-Ratio	P
Between	6397.0023	3	2132.3341		
A	5616.9020	1	5616.9020	21.6058	.0001
B	722.5003	1	722.5003	2.7791	.1005
AB	57.6000	1	57.6000	.2216	.6456
Within	9359	36	259.9722		
Total	15756.0023	39	404.0001		

the multiple correlation of $R^2 X_4 . X_2 X_3$ was computed. These two values were then substituted into the formula given earlier. A summary of the multiple correlations and the F values follows:

$$R^2_1 = R^2 X_4 . X_1 X_2 X_3 = .4060$$

$$R^2_2 = R^2 X_4 . X_1 X_2 = .4023$$

$$R^2_3 = R^2 X_4 . X_2 X_3 = .0495$$

$$R^2_4 = R^2 X_4 . X_1 X_3 = .3601$$

Main effect A:
$$F = \frac{(R^2_1 - R^2_3)/1}{(1-R^2_1)/36} = 21.606 \qquad p = .0001$$

Main effect B:
$$F = \frac{(R^2_1 - R^2_4)/1}{(1-R^2_1)/36} = 2.779 \qquad p = .1005$$

Interaction:
$$F = \frac{(R^2_1 - R^2_2)/1}{(1-R^2_1)/36} = .222 \qquad p = .6456$$

(Compare these results with those from the ANOVA in Tables 4 and 5.)

Notice that the F ratios are exactly the same as under the traditional analysis, and of course the conclusions would be the same. Also, additional information regarding the regression equation is available:

$$Y = 23.7X_1 + 8.5 X_2 + 2.4 X_3 + 21.2$$

TABLE 6

LINEAR CONSTRAINTS

Group	Cell	A_1	A_2	B_1	B_2
1	$A_1 B_1$	1	-1	1	-1
2	$A_1 B_2$	1	-1	-1	1
3	$A_2 B_1$	-1	1	1	-1
4	$A_2 B_2$	-1	1	-1	1

EXAMPLE PROBLEM: 3x3 LATIN SQUARE

Often when a researcher plans to use a Latin Square or some incomplete block design, he finds that there is no program available to analyze his particular design. However, there is always a good MR program available. If the researcher can use MR to analyze his data, then his choice of designs is not limited to the available computer programs.

A numerical example of a 3x3 Latin Square experiment having one observation per cell will be used to illustrate the computational procedure. A Latin Square of this type can be thought of as a fractional replication of a 3x3 factorial experiment. A basic assumption of this design is that the interactions are negligible.

Suppose that an experimenter is interested in

TABLE 7

DATA FOR COMPUTER CARDS

Group	Subject	X_1	X_2	X_3	X_4
1	1	1.	1.	1.	23.
	.				
	.				
	10	1.	1.	1.	67.
2	1	1.	-1.	-1.	30.
	.				
	.				
	10	1.	-1.	-1.	48.
3	1	-1.	1.	-1.	42.
	.				
	.				
	10	-1.	1.	-1.	12.
4	1	-1.	-1.	1.	31.
	.				
	.				
	10	-1.	-1.	1.	16.

TABLE 8

NOTATION MATRIX

	b_1	b_2	b_3
a_1	c_3	c_2	c_1
a_2	c_1	c_3	c_2
a_3	c_2	c_1	c_3

evaluating the relative effectiveness of three different schools (factor A) under three different methods (factor B) on three different ability groupings of students (factor C). This type of design is usually used when there are not enough Ss available for a full factorial design or the number required for a complete factorial is impractical.

Based on the results of this experiment (see Tables 9 and 10), one can conclude that there is no significant difference for any of the factors.

This same problem can be solved by using the MR technique. Table 11 shows the proper linear constraints which were based on membership or non membership in each of the three factors.

The logic of the constraint is exactly the same as for a 3x3x3 factorial experiment but with no interaction terms present.

The following multiple correlations were needed for the computations.

$$R^2_1 = R^2_{Y . (1-6)} = .2274$$

$$R^2_2 = R^2_{Y . (3-6)} = .0375$$

TABLE 9

OBSERVED SCORES

	B_1	B_2	B_3
A_1	14	11	22
A_2	8	20	13
A_3	17	6	7

* The computational procedures used are shown in Winer (4:526).

TABLE 10

3 WAY LATIN SQUARE SOURCE TABLE

Source	SS	df	MS	F
A	49.56	2	24.78	.246
B	4.22	2	2.11	.021
C	5.56	2	2.78	.028
Error	201.55	2	100.78	
Total	260.89	8		

TABLE 11

LINEAR CONSTRAINTS FOR LATIN SQUARE

Identity Vector V	Schools X_1	X_2	Methods X_3	X_4	Ability Levels X_5	X_6	Observed Scores Y
1	1	0	1	0	1	1	14
1	1	0	0	1	0	1	11
1	1	0	0	0	1	0	22
1	0	1	1	0	1	0	8
1	0	1	0	1	0	0	20
1	0	1	0	0	0	1	13
1	0	0	1	0	0	1	17
1	0	0	0	1	1	0	6
1	0	0	0	0	0	0	7

$$R^2_3 = R^2_{Y . (1-4)} = .2061$$

$$R^2_4 = R^2_{Y . (1-2)(5-6)} = .2112$$

Main effect A:

$$F = \frac{(R^2_1 - R^2_2)/2}{(1 - R^2_1)/2} = .246$$

Main effect B:

$$F = \frac{(R^2_1 - R^2_4)/2}{(1 - R^2_1)/2} = .021$$

Main effect C:

$$F = \frac{(R^2_1 - R^2_3)/2}{(1 - R^2_1)/2} = .028$$

These are identical to the results obtained from the conventional procedure. Note that if each cell represents more than one observation an entry into the data matrix for each S must be made. The number of linearly independent vectors in R^2_1 is seven ($u + X_1 + ... + x_6$) and in R^2_2 is five ($u + x_3 ... . x_6$). Therefore, df_1 for the main effect of A = 7-5=2. Since df_2 equals N minus the number of linearly independent vectors in R_1^2, df_2 = 9-7=2. The df for main effects of B and C are similarly computed.

LINEAR CONSTRAINTS

Since one of the purposes of this paper is to help the reader with the difficult task of setting up the proper linear constraints, the constraints for a 2x 3 and a 2x2x2 factorial problem are given in Tables 12 and 13. It is suggested that the serious reader may want to try his hand at setting up these restraints and use the following examples as a self

TABLE 12

LINEAR CONSTRAINTS FOR A 2x3 FACTORIAL DESIGN*

Cell	X_1	X_2	X_3	X_4	X_5	X_6
$A_1 B_1$	1	1	0	1	0	
$A_1 B_2$	1	0	1	0	1	
$A_1 B_3$	1	0	0	0	0	
$A_2 B_1$	0	1	0	0	1	
$A_2 B_2$	0	0	1	1	0	
$A_2 B_3$	0	0	0	1	1	

*X_1 represents factor A. X_2 and X_3 represent factor B. X_4 and X_5 represent the interaction term. X_6 is left for the criterion score. Note that X_3 is the product of $X_1 \cdot X_2$ and X_4 is the product of $X_1 \cdot X_3$.

check. Note that the number of variables needed for any level corresponds with the degrees of freedom for that level.

SUMMARY

In general, the assumptions underlying the use of MR are identical to the ones justifying the use of conventional procedures. MR assumes homogeneity of variance and that the Y distribution is normal. However, there can be marked deviation from these assumptions without seriously affecting the results as long as N is fairly large.

TABLE 13

LINEAR CONSTRAINTS FOR A 2x2x2 FACTORIAL DESIGN

Cell	A	B	C	AB	AC	BC	ABC Criteria
$A_1 B_1 C_1$	1	1	1	1	1	1	1
$A_1 B_1 C_1$	1	0	1	0	1	0	0
$A_1 B_2 C_1$	1	1	0	1	0	0	0
$A_1 B_1 C_2$	1	0	0	0	0	1	1
$A_1 B_2 C_2$	0	1	1	0	0	1	0
$A_2 B_1 C_1$	0	0	1	1	1	0	1
$A_2 B_2 C_1$	0	1	0	0	0	0	1
$A_2 B_1 C_1$	0	0	0	1	1	1	0

So far it has been shown that MR is equivalent to several conventional techniques. Why should one go to the bother of learning MR? It is the opinion of the authors that the MR technique has the following advantages:

1. The regression values are provided along with the F ratios, which will allow for more meaningful interpretation.

2. If there are other independent variables of interest it is easy to check their effect on Y. This is not necessarily true under conventional procedures.

3. MR is a very flexible system that can simulate most models. Often there are only a limited number of programs available and it is difficult to find one to do a specific conventional test. MR is particularly useful for analyzing unusual patterns such as those associated with many quasi-experimental designs.

4. If one is hypothesis hunting for further research, MR is an efficient way to search. The Veldman (3) Program Regran allows the use of one hundred variables, fifty regression equations and an unlimited number of F-tests.

5. Under ANOVA techniques one usually tests interactions because they are part of the model and not because of some thought out rationale. The use of MR requires more careful formulation of hypotheses.

6. MR allows the researcher to use both qualitative and quantitative independent variables.

Most classical solutions are really just special cases of a general MR analysis. It is the hope of the authors that this paper will encourage the reader to experiment with MR analysis and decide if MR has some exciting possibilities for his particular research interests.

REFERENCES

1. Bottenberg, R. A.; Ward, J. H. Jr., Applied Multiple Linear Regression, PRL-TOR-63-6, Lackland Air Force Base, Texas, 1963.

2. Cohen, J., " Multiple Regression as a General Data-Analytic System," Psychological Bulletin, 6: 426-443, 1968.

3. Veldman, D. J., Fortran Programming for the Behavioral Sciences, Holt, Rinehart, and Winston, New York, 1967.

4. Winer, B. J., Statistical Principles in Experimental Design, McGraw-Hill, New York, 1962.

A FACTOR ANALYSIS OF EPPS SCALES, ABILITY,

AND ACHIEVEMENT MEASURES [1]

PAUL W. DIXON and NOBUKO K. FUKUDA
University of Hawaii, Hilo Campus

ANNE E. BERENS
University of Hawaii

ABSTRACT

Data from 169 high school students were rotated by means of an oblique factor rotation so that from a 27x 27 variance-covariance matrix six factors were obtained. These factors were labeled: intellectual introversion, dependence, superego strength, ego-strength, independent orientation, and verbal aggression. The results suggested that the factors of intellectual introversion, dependence, superego strength, and ego-strength had concurrent validity. It was concluded that a factor analysis of the Edwards Personal Preference Schedule (EPPS), though questionable mathematically due to the fact that the EPPS is an ipsative measure, extracted psychologically relevant dimensions. The findings also suggested that introverted, socially controlled, mathematically apt students were not competing successfully with verbal, high ego-strength students in high school academic work.

EDWARDS (6) presents evidence that the scales of his test are relatively independent. This finding was substantiated by Allen (1), who found essentially the same matrix of nonsignificant relationships with the exception that the scales Affiliation vs Nurturance, Order vs Endurance, and Order-Deference showed significant positive relationships, while Deference vs Autonomy and Succorance vs Intraception showed significant negative relationships. Allen (1) suggests that these correlations show that items relevant to each of these eight variables (Order and Deference appear in two pairs) are, in a direct or increasing way, related to each other through an underlying unity along a personality continuum for each of the five pairs.

A factor analytic approach using R methodology to reveal the simple structure underlying the fifteen scales of the EPPS has normally been considered to be unacceptable by Cattell (3) and Guilford (7). They point out that the forced choice or paired comparison method of item choice used in the EPPS makes it resemble an ipsative rather than a normative measure. Guilford states that intercorrelations of ipsative measures over people using R methodology would, therefore, be improper. He also indicates that correlating ipsative scores with normative

measures leads to correlations which are difficult to interpret. Stoltz (13) cautions further that the forced choice format may also tend to make normally covarying responses diverge, due to the binary nature of the choice. Choice scores are therefore not independent estimates of the implicit probabilities of a response, but are dependent on each other, since no choice can be made without it affecting the possibility of another choice being made.

Stoltz (13) points out another difficulty in factor analyzing using R methodology. The factor of social desirability should be controlled, which he suggests could be done by partial correlation methods. He also suggests that a routine factor analysis be attempted in which the K scales from the Minnesota Multiphasic Personality Inventory (MMPI), which both Edwards and Stoltz consider to be an adequate measure of social desirability, are included to maximize the extraction of variance due to the social desirability factor, should it occur.

A factor analysis using an oblique rotation after Digman (4) was performed which included the fifteen EPPS scales, verbal and quantitive scores on the SCAT V and Q, seven teachers' ratings of students' performance, and their rank in graduating

TABLE 1

TEACHERS' RATING SCALE

	0 10 20	30 40 50	60 70	80 90 100
ACCURACY (ACCU)	Work is poor. Makes frequent errors.	Work is inaccurate and below standard.	Work is well done and reasonably accurate.	Work is of highest quality.
COOPERATION (COOP)	Disagreeable. Cannot or will not work with others.	Works with others sometimes but has difficulty.	Usually agreeable. Generally willing to help.	Always agreeable. Willing to do extra favors.
EFFORT-INDUSTRY (E-I)	Does as little as possible. Lazy.	Seldom completes required work.	Usually does work that is required. Occasionally does extra work.	Very industrious. Does extra work gladly.
INITIATIVE-LEADERSHIP (I-L)	Acts only under direction.	Seldom originates any work. Follows others.	Plans many of his activities and still needs supervision.	Marked ability to think for himself.
RELIABILITY-RESPONSIBILITY (R-R)	Neglects promise and obligations. Unreliable.	Reliable on some occasions. Often needs supervision.	Usually dependable. Conscientious.	Thoroughly dependable.
PROMPTNESS-PUNCTUALITY (P-P)	Undependable. Almost always late.	Frequently late.	Usually on time but occasionally late.	Always on time.
SELF-CONFIDENCE (S-F)	Timid. Hesitant. Easily influenced.	Appears to be over self-conscious.	Wholesomely self-confident.	Shows superb self-assurance.

class from high school. Teachers' ratings, SCAT V and Q, and rank in class (rank) are normative measures. Teachers' ratings of accuracy (ACCU) and effort and industry (E-I) have been shown by Dixon, Fukuda, and Berens (5) to be highly predictive of students' performance in high school and can be interpreted as measures of conformity to teachers' expectations in the high school setting. These ratings should therefore intercorrelate highly with those measures from the EPPS which maximize the dimension of social desirability.

The statistical technique of multiple linear regression, following Bottenberg and Ward (2), was used to assess the amount of independent contribution each scale of the EPPS, SCAT V and Q, and teachers' ratings had in prediction of rank. A similar analysis including rank was performed with all variables testing their individual contribution to prediction of SCAT V and also SCAT Q. Thus, a check on the results of the factor analysis by means of multiple linear regression was made to answer some of the questions concerning the factor pattern revealed by the oblique rotation of all variables.

The analysis using the statistical technique of multiple linear regression essentially tests the predictive validity of the oblique factor rotation and its relationship to a statistical technique yielding F ratios. It is thus an experimental test of the oblique rotation where not all the assumptions implicit in its mathematical derivation are met. Logical

criteria rather than mathematical ones were therefore used to assess the conformity of the findings to psychological assumptions.

METHOD

Subjects

The students were members of a graduating class of a large high school in Hawaii. Those selected for study had complete records on measures of SCAT V and Q, teachers' ratings, and rank. They belonged to five different English classes all taught by the same teacher. The classes were homogeneously grouped and were loosely rank ordered on SCAT V and Q and teachers' recommendations. The most academically gifted students were placed in class 1 and the least gifted in class 5; the rest were distributed in classes ranked between 1 and 5. The teacher administered the EPPS to her students throughout the course of 2 school days. One student failed to complete the EPPS, which left 169 students for the analysis.

Instrumentation

Records for the graduating class were examined and SCAT scores, rank, and teachers' ratings were recorded. The typical student had been rated by five teachers, resulting in five ratings per scale with the range being three to six ratings (see Table 1). The rating system thus conformed to the criteria given by

TABLE 3

LOADINGS ON SIX FACTORS OF EPPS SCALES, IQ CLASS, TEACHERS' RATINGS, AND R

	Intellectual Introversion	Dependence	Superego Strength	Independent Orientation	Ego Strength	Verbal Aggression
1 = n ach	0.23	0.07	0.14	0.08	0.56	0.09
2 = n def	0.03	0.10	0.86	-0.14	0.08	-0.02
3 = n ord	0.02	-0.03	0.80	0.08	-0.03	0.16
4 = n exh	-0.22	-0.15	-0.17	-0.11	0.74	-0.24
5 = n aut	-0.10	0.03	-0.10	0.68	0.17	0.14
6 = n aff	-0.37*	0.05	0.00	-0.01	0.16	-0.79
7 = n int	0.31*	0.37*	0.36*	0.12	-0.20	-0.16
8 = n suc	-0.06	0.82	-0.09	0.02	0.13	0.07
9 = n dom	-0.08	0.07	0.14	0.06	0.69*	0.02
10 = n aba	0.07	0.22	0.46*	0.02	-0.35*	-0.15
11 = n nur	-0.19	0.33	0.03	-0.09	-0.14	-0.48*
12 = a chg	0.37*	-0.31*	0.29	0.74	-0.07	-0.28
13 = n end	0.05	-0.05	0.86*	-0.09	0.31*	-0.09
14 = n het	-.40*	0.29	-0.22	0.68*	-0.09	0.03
15 = n agg	-0.02	0.24	0.06	0.08	0.12	0.61
16 = con	0.37*	0.07	-0.09	0.09	0.05	-0.45*
17 = IQ Class	-0.52*	-0.12	0.29	-0.01	-0.14	0.06
18 = SCAT V	0.48*	0.08	-0.36*	-0.06	0.16	-0.10
19 = SCAT Q	0.54*	0.16	-0.21	0.02	0.13	-0.08
20 = ACC	0.80*	0.06	-0.01	0.01	0.02	-0.03
21 = COOP	0.86*	-0.06	0.56*	0.01	-0.05	-0.02
22 = E-I	0.87*	-0.02	0.02	-0.03	-0.07	0.01
23 = I-L	0.81*	-0.01	-0.04	0.01	0.01	0.01
24 = R-R	0.86*	0.01	0.08	-0.01	-0.02	0.02
25 = P-P	0.88*	-0.02	0.05	-0.06	-0.07	0.04
26 = S-F	0.75*	-0.05	0.01	0.05	0.07	-0.01
27 = R	-0.75*	-0.08	0.15	0.05	-0.02	-0.01

*Meets criterion of $\pm.30$.

Rugg (12), who recommends the use of at least three judges and approaches the number set by Symonds (14), who recommends the use of eight judges. The ratings made by the teachers were not confined to the intervals of the scale, such as 0, 10, or 20, but were placed at any point which the teachers considered appropriate. The teachers placed checks at the point they considered appropriate for each student. The

TABLE 2

INTERCORRELATIONS BETWEEN EPPS SCALES, SCAT V AND Q, TEACHERS' RATINGS, AND R

	1	2	3	4	5	6	7	8	9	10	11	12	13
1=n ach	1.00	-0.17	-0.11	0.32	0.13	-0.29	-0.22	0.05	0.35	-0.37	-0.43	-0.04	-0.05
2=n def		1.00	0.32	-0.10	-0.24	-0.04	0.10	-0.14	-0.15	0.19	0.06	-0.10	0.21
3=n ord			1.00	-0.25	-0.18	-0.17	-0.05	-0.17	-0.25	0.10	-0.01	0.35	-0.20
4=n exh				1.00	0.10	-0.07	-0.17	-0.02	0.32	-0.30	-0.13	0.04	-0.19
5=n aut					1.00	-0.20	-0.20	-0.05	0.14	-0.25	-0.31	0.07	-0.23
6=n aff						1.00	0.06	-0.03	-0.25	0.19	0.51	0.05	-0.10
7=n int							1.00	-0.06	-0.27	0.24	0.24	0.00	-0.08
8=n suc								1.00	-0.02	-0.10	0.20	-0.18	-0.26
9=n dom									1.00	-0.37	-0.31	-0.12	0.00
10=n aba										1.00	0.29	-0.03	0.17
11=n nur											1.00	-0.08	-0.10
12=n chg												1.00	0.00
13=n end													1.00
14=n het													
15=n agg													
16=n con													
17=IQ class													
18=SCAT V													
19=SCAT Q													
20=ACCU													
21=COOP													
22=E-I													
23=I-L													
24=R-R													
25=P-P													
26=S-F													
27=Rank													

NOTE: Negative correlations between IQ Class (17) and Rank (27) and other scores are the result of smaller numbers on variables 17 and 27 showing higher placement in a homogeneously grouped class and higher rank in class, respectively.

14	15	16	17	18	19	20	21	22	23	24	25	26	27
-0.16	0.15	0.00	-0.39	0.41	0.31	0.33	0.26	0.28	0.32	0.30	0.31	0.26	-0.37
-0.23	-0.07	-0.14	0.31	-0.28	-0.24	-0.16	-0.03	-0.06	-0.12	-0.08	-0.03	-0.14	0.17
-0.20	-0.08	-0.17	0.31	-0.34	-0.26	-0.20	-0.15	-0.15	-0.23	-0.17	-0.13	-0.17	0.24
-0.05	0.13	-0.01	-0.14	0.13	0.11	0.08	0.09	0.04	0.15	0.05	0.07	0.18	-0.10
0.13	0.17	0.01	-0.10	0.15	0.09	0.07	-0.01	0.03	0.06	0.04	0.00	0.04	-0.08
-0.06	-0.48	0.18	0.10	-0.03	-0.02	-0.04	-0.04	-0.05	-0.10	-0.12	-0.09	-0.09	0.08
-0.16	-0.30	0.05	-0.14	0.11	-0.04	0.15	0.22	0.23	0.19	0.21	0.18	0.10	-0.13
0.08	0.08	0.04	-0.16	0.15	0.20	0.10	0.03	0.09	0.09	0.06	0.06	0.00	-0.16
-0.10	0.25	-0.10	-0.19	0.22	0.15	0.11	0.11	0.10	0.14	0.10	0.07	0.18	-0.17
-0.16	-0.12	0.14	0.23	-0.21	-0.16	-0.06	-0.04	-0.03	-0.15	-0.04	-0.02	-0.14	0.10
-0.13	-0.24	-.15	0.09	-0.09	-0.08	-0.06	-0.07	-0.06	-0.10	-0.06	-0.02	-0.09	0.08
-0.01	-0.09	0.14	-0.16	0.11	0.05	0.18	0.23	0.23	0.20	0.24	0.22	0.15	-0.20
-0.30	-0.02	-0.09	0.14	-0.17	-0.04	-0.07	-0.03	-0.06	-0.10	-0.06	-0.01	-0.12	0.07
1.00	0.04	-0.02	-0.01	0.02	-0.01	-0.18	-0.20	-0.19	-0.13	-0.20	-0.21	-0.12	0.11
	1.00	-0.19	-0.02	-0.05	-0.09	-0.06	-0.09	-0.09	-0.04	-0.04	-0.05	0.04	0.08
		1.00	-0.33	0.30	0.39	0.36	0.39	0.34	0.33	0.34	0.35	0.24	-0.35
			1.00	-0.83	-0.62	-0.69	-0.61	-0.60	-0.67	-0.60	-0.56	-0.55	0.73
				1.00	0.65	0.67	0.58	0.56	0.63	0.57	0.54	0.52	-0.73
					1.00	0.67	0.53	0.55	0.58	0.57	0.55	0.41	-0.68
						1.00	0.86	0.92	0.90	0.90	0.84	0.75	-0.87
							1.00	0.91	0.90	0.94	0.87	0.80	-0.75
								1.00	0.90	0.93	0.86	0.75	-0.82
									1.00	0.90	0.82	0.84	-0.80
										1.00	0.91	0.79	-0.80
											1.00	0.70	-0.73
												1.00	-0.65
													1.00

TABLE 4

INTERCORRELATIONS AMONG FACTORS

Factor	Intellectual Introversion	Dependence	Superego Strength	Independent Orientation	Ego Strength	Verbal Aggression
1	1.00	0.17	-0.30	0.16	0.33	-0.22
2		1.00	-0.11	0.25	0.05	-0.31
3			1.00	-0.17	-0.39	-0.04
4				1.00	0.36	0.09
5					1.00	0.42
6						1.00

SCAT was administered in the fall of the students' senior year in high school. Rank in class was based on the first five semesters of work in high school.

Procedure

The marks placed by the teachers on the rating scale of each student were measured from the zero point on the scale in centimeters to one place accuracy. Thus, for the scale of accuracy a student might have scores of 9.6, 9.6, 9.8, 8.5, and 8.9, as measured in centimeters. For each student, eleven different school related scores were obtained: IQ class, Verbal (SCAT V), Quantitative (SCAT Q), averages of the teachers' ratings on the scales of Accuracy (ACCU), Cooperation (COOP), Effort and Industry (E-I), Initiative and Leadership (I-L), Reliability and Responsibility (R-R), Promptness and Punctuality (P-P), Self-Confidence (S-F), and ranks. Scores for the fifteen EPPS scales, including the consistency scale, were obtained for each student. These were need for achievement (n ach), deference (n def), order (n ord), autonomy (n aut), affiliation (n aff), intraception (n int), succorance (n suc), dominance (n dom), abasement (n aba), nurturance (n nur), change (n chg), endurance (n end), heterosexuality (n het), aggression (n agg), and consistency (con, a measure of reliability of response).

RESULTS

The subroutines (PERSUB) of Bottenberg and Ward (2) and an IBM 360-50 computer were used to determine the extent to which post-high school destination and sex variables contributed to differences in SCAT, teachers' ratings, and rank. PERSUB was used to perform the statistical technique of multiple linear regression to determine F ratios and exact probability values to 4-place accuracy. Multiple linear regression using PERSUB presents an efficient method for programming computation of the estimated probability for any specific F value. The estimate is based on the actual distribution of scores and may be used with degrees of freedom ranging from 4 to 1,000 (11). The F ratio is computed between a full model regression equation containing all predictor variables under consideration and a restricted model.

In the restricted model in this analysis, the information about a particular variable is not included in the restricted model's equation, and the predictive efficiency of this equation is compared with the predictive efficiency of the full model's equation, in which all the information for each variable is included. For example, in predicting IQ scores the full model would include the male-female differences in the population, while the restricted model omits the information that male-female differences existed. If knowledge of male-female differences helped in the prediction of IQ scores, there would be a significant difference using the F ratio statistic between the equation containing male-female differences and the restricted model, which does not include this information.

The variables were analyzed by means of image analysis (8), which may be viewed as an alternative to factor analysis. In image analysis, the analysis is performed not on the scores, but on the "images" of the scores, i.e., the parts of the scores which are predictable—in a multiple regression sense—from the n-1 other variables in the battery. From such image scores a variance-covariance matrix may be formed and factored in much the way in which a correlation matrix may be factored. However, as compared with the Thurstone model of factor analysis, image analysis has certain distinct advantages: (1) There are no communalities to be estimated, as it is the shared variance—in the sense of a squared multiple correlation—which is the variance to be accounted for. (2) Unlike the Thurstone model, there are no imaginary factors, i.e., the matrix is a Gramian matrix. Although image analysis is conceptually different from factor analysis, it is, however, quite unlikely that major differences in interpretation would result from its use in place of the more traditional methods.

In the present study, the variance-covariance matrix of image scores was reduced by the principal-axes method to six factors. These were then transformed ("rotated") by Digman's (4) variant of the Harris-Kaiser (9) method of oblique transformation, a technique which appears to represent a substantial advance in the field of "factor rotation."

The correlations among the twenty-seven variables are shown in Table 2; six factors were found to have an eigenvalue >1.00. Following the rule of Kaiser (10), the solution was obtained by rotating the factors with eigenvalues greater than 1. The variance-covariance matrix was rotated by the Varimax procedure with the results shown in Table 3. No allowance was made for differences between the sexes in this analysis due to the small number of students in the sample.

Table 4 shows the intercorrelations (factor cosines) among the six factors derived from the factor analysis. Table 4 shows, using as criterion a factor correlation of $\pm.30$, that Factor 1 is positively related to Factors 3 and 5, Factor 2 is negatively related to Factor 6, Factors 3 and 5 are negatively related, Factors 4 and 5 and Factors 5 and 6 are positively related.

The statistical technique of multiple linear regression was used with a full model of all variables predicting to the criteria of interest. Each variable was omitted sequentially but one at a time to assess the amount of individual contribution in prediction of the criterion.

The first criterion used for prediction was rank, which was considered important because it was a measure of academic achievement in a high school setting. SCAT V, COOP, and E-I showed significant prediction of rank ($p<.05$). Teachers' ratings of ACCU showed significant independent contribution in prediction of rank ($p<.01$).

A further test was made with all variables predicting to the criterion, SCAT V, as shown in Table 6. This was considered important since the SCAT are widely used and readily available from the Educational Testing Service Cooperative Test Division, Princeton, New Jersey. Those variables showing significant independent contribution to SCAT V were SCAT Q and rank ($p<.05$). Those showing significant prediction at the .01 level were n ach, E-I, and, obviously, IQ class.

In prediction of SCAT Q the variables n int, E-I, and SCAT Q made significant independent contributions ($p<.05$). In addition to these variables, teachers' ratings of ACCU and S-F made significant independent contributions to prediction of SCAT Q at the .01 level of significance.

DISCUSSION

The results of the 6-factor rotation to an oblique solution (see Table 3) shows that for the first factor the following variables meet the criteria of $\pm.30$ (the criterion $\pm.30$ was chosen in order that each variable would load on at least one factor): affiliation (the need to form friendships) is negatively related, intraception (the need to observe and analyze one's own and other's feelings and motives) is positively related, change (the need for novelty in everyday things) is positively related, heterosexuality (the need to have social and physical relationships with members of the opposite sex) is negatively related, all the teachers' ratings, SCAT V and Q, R and IQ class are positively related (the last two variables have negative signs, since smaller scale values are higher rankings). The consistency or reliability

measure of the EPPS weighs positively on this factor also.

Since teachers' ratings of ACCU and R-R can be said to be measures of conformity to teachers' expectations in the school system, they may also be said to be variables that maximize social desirability. Factor 1 is therefore a reliable factor which extracts the major part of variance due to the presence of social desirability in the EPPS. Considering Factor 1 as a scale, it may be said to be measuring intellectual introversion within a high school setting. Social conformity, as a tendency to answer items in a socially desirable way, is positively related to success in high school according to this analysis.

Factor 2 was labeled the dependency factor, since succorance (the need to seek encouragement, help, and affection from others) showed the highest loading, intraception loaded positively as did nurturance (the need to help friends and others less fortunate), and change met criterion with a negative weighting. Intraception is shown to be a complex variable, since it shows positive loadings on this factor as well as on Factors 1 and 3. The reliability of this factor cannot be gauged by the factor loading of the consistency measure of the EPPS, which is negligible for this factor: it must be judged in relation to the other factors in regard to its psychological relevance. The individual with a high score on a scale made from the items meeting criterion on Factor 2 would be a dependent, kindly, introverted person with a tendency to support the status quo.

Factor 3 shows the highest loading for deference (a need to find out what others think), second highest for endurance (need for job completion), order (need for neatness and planfullness), abasement (self-punishment for perceived guilt), intraception, and a negative loading for SCAT V. This factor could be labeled the strong superego factor. An individual with a high score on a scale from the items meeting criterion on Factor 3 would be retiring, self-effacing, and nonverbal. Table 4, which gives the correlations among factors, reveals that Factors 1 and 3 are negatively related (criterion $\pm.30$). Thus, the intellectual introvert is typically not the individual with powerful superego development. Consistency does not reach criterion on Factor 3, however, so it must be judged on psychological rather than statistical grounds.

Factor 4 has, as its highest loading, autonomy (the need for independence and freedom from constraint). The next highest loading is heterosexuality (the need to have social and physical relations with members of the opposite sex) and change. High scorers from a scale derived from Factor 3 would have a need to be independent from societal sanctions. A freedom to express heterosexual interest would result from their subjective independence from societal constraints.

Factor 5 has for its highest loading the variable exhibition (the need to be noticed and to have one's personal achievements talked about), the next factor loading weights were dominance (the need to be a leader and to influence others), achievement (the need to be successful), and endurance with a negative loading for the need for abasement. The individual

TABLE 5

THE INDIVIDUAL CONTRIBUTION OF EPPS SCALES, TEACHERS' RATINGS, IQ CLASS, AND SCAT V AND Q IN PREDICTION TO R

Source	R^2	F
Full R	0.8304	
Achievement	0.8292	1.02
Deference	0.8285	1.59
Order	0.8304	0.02
Exhibition	0.8303	0.09
Autonomy	0.8302	0.12
Affiliation	0.8299	0.40
Intraception	0.8304	0.01
Succorance	0.8380	1.98
Dominance	0.8284	1.62
Abasement	0.8291	1.09
Nurturance	0.8298	0.50
Change	0.8277	2.23
Endurance	0.8289	1.28
Heterosexuality	0.8297	0.58
Aggression	0.8283	1.77
IQ Class	0.8265	3.40
SCAT V	0.8238	5.49*
SCAT Q	0.8274	2.47
ACCU	0.8169	11.32**
COOP	0.8254	4.16*
E-I	0.8241	5.23
I-L	0.3801	0.26
R-R	0.8288	1.34
P-P	0.8287	1.42
Self-Con	0.8304	0.01

NOTE: Full model regression equation: unit vector + Edwards scales + IQ class + SCAT V and Q + teachers' ratings predicting to R.

*$p < .05$ level of significance.

**$p < .01$ level of significance.

df=1/142

vidual choosing items which would score positively on a scale made from this factor would be high in ego-strength.

Table 4 indicates that the pairs of Factors 1 and 5, and 4 and 5 are positively related. Thus, intellectual introversion, independent orientation or outlook, and ego-strength form a constellation of personality factors which relate positively to one another. The sixth factor also relates positively to Factor 5.

The variables meeting criterion for Factor 6 are, in order of importance, affiliation, negative loading; aggression (the need to attack, to argue, to become angry) with a positive loading; nurturance, negative loading; and consistency, negative loading. If consistency is a measure of reliability and not a personality variable, which might be construed from the variable's loading on Factor 6, then the reliability of Factor 6 is probably very poor and its meaning doubtful. The factor of social acceptance or conformity (Factor 2) relates negatively to Factor 6. A negative relation between the expression of hostility and aggressive impulses and social control would be expected from a psychological point of view. [2]

Intellectual introversion is related most closely to performance in high school, but measures of ego-strength also relate to academic performance. Individuals rating high on ego-strength would also tend to be verbally aggressive, autonomous in their outlook upon societal constraints and would tend not to have a strongly intra-punitive superego. The intellectual introvert would not tend to have a need to accept blame and guilt and be deferent. Thus, the self-analytical, high ego-strength individual, rather than the dependent and self-abasing person, would be best suited to achieve in the high school milieu. The successful high school student would probably be well-structured and planful and would do neat, accurate, and careful work.

The predictive validity of each variable was assessed using multiple linear regression to find which variables showed significant independent contribution to prediction of R. SCAT V, COOP, and E-I showed significant independent contributions to prediction of R ($p < .05$), and teachers' ratings of ACCU showed prediction at less than the .01 level of significance. Table 5 thus shows that the variables of Factor 1 do not all show significant independent contributions to prediction of R. The EPPS scale for intraception, change and heterosexuality might therefore be considered to be related to the intellectual introversion factor but is not significant in a predictive sense. Factor 1 is therefore more of a personality dimension than a scale which shows predictive validity within the scholastic environment.

The predictive validity of all variables was assessed in relation to SCAT V, as is shown in Table 6. SCAT Q and R showed significant independent contributions to SCAT V ($p < .05$). Achievement, IQ class, and E-I also showed prediction of SCAT V at less than the .01 level of significance. Those scoring well on SCAT V would thus be in a more selective, heterogeneously grouped class, show a high need for achievement, be rated as industrious in school, rank high in their graduating class from high

school, and score relatively well on SCAT Q. Need for achievement is related to Factor 5 (ego-strength), which is also related to the verbal abilities factor, as is indicated above. The EPPS scale of need for achievement and probably Factor 5 of this factor analysis might therefore be said to have some substantiation of their importance in a high school setting from the test of predictive validity. Ego-strength and verbal ability are probably important factors in prediction of success in high school.

A further prediction was made to SCAT Q using all other variables. Intraception, SCAT V, and E-I showed significant prediction of SCAT Q (p<.05). ACCU showed significant independent contribution to prediction of SCAT Q also (p<.01). (It should be noted that R did not make significant independent contribution to SCAT Q.) A student scoring well on SCAT Q would then be accurate, neat, industrious, and introverted. Thus, the high SCAT Q student would differ on salient personality dimensions from the high SCAT V student, according to this analysis, even though SCAT V and Q show significant prediction of each other. The relationship between need for intraception and Factor 1 is shown to be in relation to the quantitative dimension of intellectual functioning which is part of Factor 1. Intraception, the self and other analytical qualities revealed by this scale, in addition to its presence in Factor 1, is factorially complex, since it is also part of the social conformity and superego-strength factors. The significant prediction of SCAT Q by intraception may thus be said to partially validate the loading of intraception on Factor 1. By contrast, need for achievement is related to SCAT V and Factor 5 (ego-strength), and SCAT V is predictive of R as well as being part of Factor 1. The ambitious, verbal student might well be said to be at an advantage in a high school milieu, while the controlled, quiet, introverted, mathematically apt student is not. The academic potentials of the latter student may well go unrecognized and unrewarded in an academic milieu of this nature. Counseling and selection of students from high school for college work should take into account that an important segment of the academically gifted students in a high school population, to some degree, go unrecognized as compared with their highly verbal, achievement-oriented peers.

SUMMARY AND CONCLUSION

The EPPS was administered to 169 students in five homogeneously grouped classes varying from high to low ability. Measures from a teachers' rating schedule of classroom performance, SCAT V and Q and rank in graduating class were obtained from the school records.

An oblique rotation was performed on all twenty-seven variables using R methodology to obtain the simple structure. This revealed six factors which were labeled: (1) intellectual introversion, (2) dependence, (3) superego strength, (4) independent orientation, (5) ego-strength, and (6) verbal aggression. It was found that intellectual introversion was negatively related to superego strength and positively to ego strength; dependence was negatively related to verbal aggression; superego strength and ego-strength were negatively related; while the parts

TABLE 6

THE INDIVIDUAL CONTRIBUTION OF EPPS SCALES, TEACHERS' RATINGS, IQ CLASS, R, AND SCAT Q IN PREDICTION TO SCAT V

Source	R^2	F
Full V	0.7675	
Achievement	0.7568	6.56**
Deference	0.7667	0.46
Order	0.7674	0.04
Exhibition	0.7659	0.99
Autonomy	0.7625	3.04
Affiliation	0.7654	1.29
Intraception	0.7628	2.89
Succorance	0.7671	0.25
Dominance	0.7629	2.81
Abasement	0.7671	0.26
Nurturance	0.7666	0.56
Change	0.7673	0.10
Endurance	0.7673	0.10
Heterosexuality	0.7635	2.43
Aggression	0.7684	0.00
IQ Class	0.6908	46.84**
SCAT Q	0.7594	4.94*
ACCU	0.7656	1.17
COOP	0.7642	2.04
E-I	0.7569	6.45
I-L	0.7674	0.07
R-R	0.7673	0.09
P-P	0.7675	0.01
Self-Con	0.7673	0.10
Rank	0.7585	5.50*

NOTE: Full model regression equation: unit vector = 15 Edwards scales + IQ class + SCAT Q + teachers' ratings + R predicting to SCAT V.
*p<.05 level of significance.

**p<.01 level of significance.

df=1/142

TABLE 7

THE INDIVIDUAL CONTRIBUTION OF EPPS
SCALES, TEACHERS' RATINGS, IQ CLASS, R,
AND SCAT V IN PREDICTION TO SCAT Q

Source	R^2	F
Full Q	0.6223	
Achievement	0.6180	1.63
Deference	0.6202	0.79
Order	0.6205	0.66
Exhibition	0.6210	0.50
Autonomy	0.6213	0.37
Affiliation	0.6223	0.00
Intraception	0.6069	5.78*
Succorance	0.6204	0.73
Dominance	0.6221	0.09
Abasement	0.6207	0.61
Nurturance	0.6202	0.78
Change	0.6138	3.18
Endurance	0.6205	0.66
Heterosexuality	0.6223	0.00
Aggression	0.6144	2.98
IQ Class	0.6209	0.51
SCAT V	0.6091	4.96*
ACCU	0.6025	7.45**
COOP	0.6221	0.08
E-I	0.6121	3.85*
I-L	0.6198	0.93
R-R	0.6193	1.12
P-P	0.6200	0.87
Self-Con	0.6030	7.25**
Rank	0.6157	2.47

NOTE: Full model regression equation: unit vec-
tor + 15 Edwards scales + IQ class + SCAT V +
teachers' ratings + R predicting to SCAT Q.
*p<.05 level of significance.

**p<.01 level of significance.

df=1/142

of the factors independent orientation and ego-
strength, and ego-strength and verbal aggression,
were found to be positively related.

A factor analysis and a prediction study using the
EPPS, teachers' ratings of classroom behavior and
a standardized test of verbal and quantitative skills
showed that highly verbal students with greater ego
strength (need for achievement) tend to dominate in
high school grades. The mathematically apt student
who has greater superego strength (accuracy and ef-
fort and industry) and is perhaps more dependent,
does not. Counseling of students in which quantita-
tive skills are considered an important part of intel-
lectual functioning is suggested if an important seg-
ment of the intellectually gifted student body is not
to be neglected. Factor analysis of the EPPS, though
questionable from a strictly mathematical viewpoint,
does reveal a simple structure with a strong sugges-
tion of psychological relevance in a high school set-
ting.

FOOTNOTES

1. This research was supported by NSF funds ad-
   ministered by the research council of the Uni-
   versity of Hawaii. We are indebted to John M.
   Digman and Elsie H. Ahern for valuable assis-
   tance with this paper.

2. The factors obtained in this study resemble the
   Milton-Lipetz solution though item pairs were
   not eliminated in this study. They found five
   factors which were a need for interpersonal re-
   lationship and affiliation, need for hostile de-
   dependency, need for status dominance, need
   for structure and orderliness, and need for
   freedom or independence. Milton, G. A.;
   Lipetz, M.E., "The Factor Structure of Needs
   as Measured by the EPPS," Multivariate Be-
   havioral Research, 1:37-46, 1968.

REFERENCES

1. Allen, R.M., "Edwards Personal Preference
   Schedule Intercorrelations for Two Groups,"
   Psychological Record, 7:87-91, 1957.

2. Bottenberg, P.A.; Ward, J.H., Applied Multi-
   ple Linear Regression, Technical Documentary
   Report PRL-TDR-63-6, Defense Documentation
   Center, Defense Supply Agency, Washington,
   March, 1963.

3. Cattell, R.B., "Psychological Measurement:
   Ipsative, Normative, and Interactive," Psy-
   chological Review, 51:292-303, 1944.

4. Digman, J.M., "The Procrustes Class of Fac-
   tor Analytic Transformations," Multivariate
   Behavioral Research, 2:89-94, 1967.

5. Dixon, P.W.; Fukuda, Nobuko K.; Berens, Anne
   E., "Effectiveness of Teachers' Ratings, Sex,
   and SCAT Scores as Predictors of Rank in High
   School Class," The Journal of Experimental
   Education, 37:21-26, 1969.

6. Edwards, A.L., Edwards Personal Preference
   Schedule: Manual (revised), New York: Psy-
   chological Corporation, 1959.

7. Guilford, J.P., "Psychological Measurement
   125 Years Later," Psychometrika, 26:101-127,
   1961.

8. Gutman, L., "Image Theory for the Structure
   of Psychological Variates," Psychometrika,
   18:277-296, 1953.

9. Harris, D.W.; Kaiser, H.F., "Oblique Factor
   Analytic Solution," Psychometrika, 29:347-362,
   1964.

10. Kaiser, H.F., "The Application of Electronic
    Computers to Factor Analysis," Educational

and Psychological Measurement, 20:141-151,
1960.

11. Neel, G.I., Estimation of Probabilities Asso-
    ciated with the F Statistic by Digital Computer
    Techniques, Technical Documentary Report
    PRL-TDR-63-7, Defense Documentation Cen-
    ter, Defense Supply Agency, Washington,
    March, 1963.

12. Rugg, H.O., "Is the Rating of Human Charac-
    ter Practicable?" Journal of Educational Psy-
    chology, 13:30-40, 81-93, 1922.

13. Stoltz, R.E., "Note on the Intercorrelations of
    Edwards," Psychological Reports, 4:239-241,
    1958.

14. Symonds, P.M., Diagnosing Personality and
    Conduct, Appleton, 1931, p. 602.

Book Reviews

his own concepts, to be intellectually self-changing. But Borton's paradigm includes more. While classical ed-
ucation confined discovery and self-change to the cognitive domain, Borton extends it to the affective domain:
the student will learn the skills and processes through which he can change his relationship to himself and to
others. It is the inclusion of intra and interpersonal skills and processes which distinguishes Borton's approach
from the classical view.

In the first part of the book, Borton describes his earliest attempts to involve students in their education.
Like many contemporary educational innovators, Borton's teaching experiences were intimately tied to the urban-
student struggle with issues of race and identity. His efforts were directed toward legitimizing white and black
students' exploration of feelings about what it means to be white or to be black; Borton believed that for students
to get involved in school, they must begin with their own concerns. While Borton describes success in enabling
students to explore their concerns, he also describes failure in moving students from these concerns to the at-
tainment of broader educational goals. His students, although involved in personally relevant issues, were not
learning the skills of self-change.

To facilitate movement toward this broader goal, Borton presents a simple information processing model
indicating which processes or skills should be taught. The model consists of three stages: sensing, transform-
ing, and acting. According to this model, the student needs to learn the skills of experiencing his world (sens-
ing), of analyzing the meaning of his experience (transforming), and of responding in new ways based on his
analysis (acting).

This three-stage model raises some interesting theoretical questions: How do children "sense" the world
at different ages? What cognitive and affective processes are necessary for understanding one's own experience?
What skills are necessary for the behavioral application of cognitive understanding?

In regard to how these skills are taught, Borton uses three terms corresponding to each information pro-
cessing stage: "What," "So What," and "Then What." The "Whats" are experiences of the student; the "So
Whats" are the analyses and cognifications of the experiences; the "Then Whats" are the action implications fol-
lowing from the analyses. These actions, when carried out, provide new data for the information processing cy-
cle. Thus learning proceeds from experience to cognification to application to new experience. It is, funda-
mentally, an inductive approach to learning.

Some important pedagogical questions emerge: How does one construct meaningful experiences for child-
ren? How can one proceed sensibly from an experience to the cognification of that experience? How can the child's
applications of his experience be monitored to ensure continued growth?

If the book suffers it is in its application of the model. For the model to come alive, teachers need meth-
ods for assessing student concerns, and they need to know the kinds of experiences which they can provide to
connect with those concerns. Borton provides a wealth of ideas for locating specific student concerns; these
ideas range from the use of role playing to the use of poetry. Similarly, he provides activities that can be used
to teach processes while connecting with the assessed student concerns; these include fantasy writing and sim-
ulation games. Unfortunately, his ideas appear to be more a function of his own creative spark than to be meth-
ods which other teachers can systematically utilize. It may be premature to expect a new paradigm to present
a complete methodology for assessment and instruction; nonetheless, the problems which remain might have
been more clearly articulated.

CHILDREN'S LITERARY SKILLS

HOWARD GARDNER and JUDITH GARDNER[1]
Harvard University

ABSTRACT

Twelve Ss at each of four age levels performed a story completion and retelling task designed to measure children's skills of understanding, retelling, and creating literature, and to test various theoretical formulations about aesthetic development. Sensitivity to literary style was also examined. Though Ss 11 or 12 years of age generally evinced the most literary skill, a few children at each age level were outstanding. Characteristic performances at each age level, individual differences, trends in the development of literary skill, and educational implications are discussed.

IN DESCRIBING children's competence in the literary realm, several competences seem worthy of examination: (1) capacity to follow the manifest plot of the story; (2) sensitivity to subtler nuances—tone, style, underlying forces and tensions; (3) ability to tell and retell stories to others; (4) skill in creating novel plots, developing themes, and/or choosing words aptly. Though a number of investigators have collected children's stories (1, 8) their analyses have been concerned chiefly with the subject matter of the stories or the relationship between the child's personality and his literary output. As a result, little is known about the extent and range of children's skills in creating, communicating, and comprehending stories. To probe these skills, the authors devised a story-completion and retelling task which could be administered to diverse age groups. Differences across and within age groups were of interest as well as indices suggesting a general trend of literary development. The relationship between literary skills and other kinds of cognitive capacities was of particular interest and the following possibilities were considered: literary skill improves gradually with age; adolescents are significantly more skilled than preadolescents at handling literary tasks, just as they are more skilled at other aesthetic assignments (2, 7); the capacity to perform logical operations impairs literary competence (5); a major spurt in literary development occurs at about age 7, as in other cognitive realms (4, 10).

METHOD

Subjects

Twelve first graders (modal age 6), twelve third graders (modal age 8), twelve sixth graders (modal ages 11-12), and twelve ninth graders (modal ages 14-15) were selected at random from two schools having a predominantly middle-class population.[2] Both sexes were equally represented in the sample.

Materials

Two plots, one original, the other adapted from an old poem, were each recast by the first author into two markedly divergent styles. Style A was reminiscent of fairy tales, with long complex sentences, remote setting, antiquated dialogue, a neutral pose toward reader and subject matter. (Once upon a time, many years ago, in a little town far across the sea, there lived a man who...) Style B, more contemporary and colloquial, featured short pungent sentences, slang, numerous exclamations, and asides by the narrator. (Here's a story about a really rich man. A guy with more bags of gold than you ever dreamed of.) Each story totaled about 350 words in Style A, 300 words in Style B; both featured good and evil characters and a central crisis.[3]

Procedure

So that each S could have the opportunity to create in both styles, a 2-phase procedure involving two experimenters (E_1 and E_2) situated in two rooms (A and B) was employed.

In phase I, S entered Room A where E_1 read to him twice a "story without an ending." S was in-

structed to attend carefully to the story and then "to make up an ending which you like and which sounds right for the story." During this phase, S heard either plot 1 or plot 2 in Style A or Style B. After S created his ending, he was told to go to Room B and to retell the story he had heard and the ending he had constructed to E_2 who "does not know the story." In order to create a set for exact repetition, S was reminded of the initial sentences of the original just before proceeding to E_2.

In phase II, the crucial parts of the initial phase were repeated in Room B. After he had related the original story and ending, S was told the second plot in the second style and asked to make up an appropriate ending. Thereafter he was instructed to return to Room A and to retell the second story and its ending to E_1. Upon completion of the story retelling to E_1 S was asked a few general questions.

This rather cumbersome procedure insured that each S heard both of the plots and both of the styles and could reveal his assimilation of them in a natural situation. The entire session was tape recorded and later transcribed. Order of story, style presentation, and experimenter were counterbalanced.

Measures

A set of new measures was devised for this exploratory study. Capacity for recall was determined by noting how many of the six major facts and the twenty-six details in each story were included in the retelling. Children received a point for each fact and 1-2 points for each detail recalled. Understanding was judged by the contents of the recall (did S get the point of the story?) and the degree to which the endings took into account the important conflicting forces in the stories. Creative ability was assessed by an originality score based on uniqueness and aptness of the endings; protocols were coded independently by the authors and given an originality score from 0 to 6. Inter-judge reliability was .97.

Cutting across these skills, and of particular interest, was the child's sensitivity to literary style (3). The six major ways in which the two styles differed were specified and Ss were independently scored on the extent to which their endings and their retellings possessed these distinctive characteristics. Protocols could receive scores from 0—24; Ss scoring at least 5 were considered possibly sensitive to style, those scoring at least 10 definitely were sensitive to style. A high score was possible only if a Ss; protocols differed significantly in style. Reliability was .94 for endings, .90 for retellings.

RESULTS

Three trends recur throughout the data analysis; the lower scores of the first graders; the clustering of the means of the three older groups, with the sixth graders generally scoring higher; the wide range of scores on most measures at each level. The slightly greater skills of the sixth graders and the wide range of scores are indicated by the summary statistics in Table 1.

Because of the clustering tendencies, two analyses of variance were performed for each of the principal measures: a one-way analysis on all four age groups, and a one-way analysis on the three older groups. Results are presented in Table 2 and drawn on in the discussion of specific findings.

Discussion of Specific Factors

Length and Nature of Endings. In this readily measured category, the findings of the overall study are nicely epitomized. First graders differed significantly from the older groups, generally adding only a few words to the stories; older children usually added at least one hundred words to the stories, covering a number of events and often attempting to balance the forces of good and evil presented at the beginning of the story. Only eight of the first graders' endings included more than one event, while at least nineteen of the endings of each of the older grade levels contained a number of events.

Recall: Nearly every S recalled most of the major facts in the story, a clear indication that they were attending to the task and possessed some understanding of the story. The three older groups were almost flawless. On the other hand, the first graders included almost no details in their retelling, while some of the older Ss mentioned nearly every detail.

Originality: Few of the first graders avoided the most banal endings. The third graders often presented endings which were lengthy and unique but not particularly appropriate. The more talented sixth graders presented the most interesting and original endings, but the overall performance of their class did not differ from the ninth graders' performance. The latter group provided endings of limited originality but sensed what was appropriate and responded well to the formal demands of the stories. Their endings resembled one another, suggesting that they knew how such stories should end.

Sensitivity to Style: Performances were generally disappointing. Few Ss of any age rendered endings which reflected the distinctive styles of the stories. According to the guidelines described above, one S at each of the three higher grade levels was judged definitely sensitive to style in his endings, while two third graders, two sixth graders, and three ninth graders were classed as possibly sensitive.

Style sensitivity was much greater in the retellings, but despite the efforts of the coders to discount specific word choice, this measure was probably confounded to some extent with memory. Differences among Ss were vast at each age level, this measure being the only one in which the first graders did not perform significantly worse than the older groups. The ability of the first graders to remember certain striking phrases and to recall the first line of the story probably contributed to the appearance of style sensitivity. Two first graders, four third graders, six sixth graders, and three ninth graders were judged sensitive to style in their retellings; two first graders, four third graders, three sixth graders, and three ninth graders were considered possibly sensitive.

TABLE 1

MEANS, STANDARD DEVIATIONS AND RANGES ON PRINCIPAL MEASURES

Statistic	Grade level			
	first	third	sixth	ninth
LENGTH IN LINES				
Mean	1.33	11.54	16.25	11.37
S.D.	.59	14.12	13.66	7.62
Range	1—3	1—54	1.5—47.5	6—31.5
MAJOR FACTS (maximum score 12)				
Mean	9.21	11.12	11.67	11.62
S.D.	1.79	.97	.26	.63
Range	6—11.5	9—12	11—12	10—12
DETAILS (maximum score 78)				
Mean	*	23.29	30.88	24.96
S.D.		9.88	10.54	12.72
Range		7.5—44	17.5—54	9.5—48.5
ORIGINALITY (maximum score 6)				
Mean	.73	2.81	3.31	3.14
S.D.	.56	1.45	1.42	.84
Range	0—2	.25—5.25	1—5.25	2.—4.25
STYLE SENSITIVITY IN ENDING (maximum score 24)				
Mean	.79	3.25	3.75	4.46
S.D.	1.05	3.01	2.80	4.15
Range	0—4	.5—10.5	.5—10	.5—15.5
STYLE SENSITIVITY IN RETELLING (maximum score 24)				
Mean	5.04	8.33	9.33	7.04
S.D.	5.08	4.70	5.66	6.79
Range	.5—16.5	2—19	1.5—19.5	0—19

* Measure inapplicable for reasons cited in footnote 3.

TABLE 2

ANALYSES OF VARIANCE OF SCORES ON PRINCIPAL MEASURES

Age levels included	F scores on various measures					
	Length	Major facts	Details	Originality	Style ending	Style retelling
4 age levels	3.90*	13.20**	——	12.40**	4.03*	1.28
df = 3, 44						
3 higher age levels	.57	2.39	1.44	.45	.36	.45
df = 2, 33						

* p. = < .025
** p = < .01
Note: No other differences significant.

Moral Stance of the Ending: The youngest Ss made no attempt to balance the scales of justice. They allowed the powerful forces to triumph completely or rewarded good forces without regard to the fate of evil. With increasing age, Ss were more likely to focus on the villian, having him repent, yield his power, or be punished. Both sixth graders and ninth graders included many reversals of fortune in the narrative, but the ninth graders more often presented the kind of moralistic ending traditional for fairy tales.

Aspects of Understanding: Though only direct questions can unambiguously establish understanding, the content of the endings and the retellings definitely indicate that every S had some understanding of the stories. That the comprehension of younger Ss was incomplete is suggested by the briefness of their endings and their failure to take into account the forces and power conflicts in the stories. While the first graders did not integrate suggestive elements of the plot in their endings, over one half of the stories of the third graders, and over three-fourths of the stories of the sixth and ninth graders included some integration of pregnant facts or details. Posttest interviews revealed that only the ninth graders were explicitly aware of the style differences in the two plots; while this awareness may signal a greater appreciation of the formal properties of literature, it does not imply superior ability in recreating that style.

DISCUSSION

Explanatory studies are designed as much to generate as to test hypotheses. Generalizations are accordingly risky, yet in a field as uncharted as literary skill, some attempt to organize and interpret findings seems justified.

The typical performance at each age level can be characterized. First graders treat the stories like a series of strips in a comic, calling for one additional line to complete the message. These closing lines are less inappropriate than incomplete, as they do not take adequate account of the various forces extant in the stories. The general lack of originality contrasts with young children's frequent imaginativeness in spontaneous storytelling. Probably the structured nature of the task restricts the youngster's creative powers. Understanding is limited, as the young Ss lack sufficient personal experience, cognitive complexity, and familiarity with literary convention.

An apt description of the third grader's stories is picaresque. Subjects interpret the task as an occasion to list a long series of events involving a hero. Often these episodes are borrowed from other stories and may be inappropriate; yet the most gifted third graders relate endings which are both inventive and relevant.

Sixth grade might be viewed as the watershed of literary development. Subjects understand the stories, select appropriate endings, and are in control of syntax and ideas. Superior performances in most skills are found at this level. If endings are cartoon-like, they are considered; if picaresque, the child maintains control of the story's drift. An increasingly cognitive orientation is evident, as characters "think," "doubt," and "bargain"; good and evil are balanced. A majority of the children combine the daring inventiveness of the younger child with the control and direction of the older child. Even the less talented seem to retain promise.

Self-consciousness and self-criticism often combine to hinder the literary productivity of the oldest group. Most are already "professional" in their approach, competent in executing the task, but less imaginative, less alert to details, and less able to preserve stylistic nuances than the sixth graders. The oldest group excels in psychological insight and in the ability to discuss literary prop-

erties; but Ss tend to rework the material into their own or their peers' way of speaking. The advent of formal operations, which enables the child to fit different contents into the same operational structure, paradoxically seems to diminish the child's ability to remain within the style, rhythm, or tone created by an author. Thus, among the various theoretical trajectories considered, the evidence favors two positions: the spurt in capacity following the ages 5-7 and the diminishing of sensitivity to language following the onset of puberty, perhaps due to the advent of formal operations (5, 6). No evidence in favor of gradual improvement or adolescent superiority was found.

Though these age differences seem genuine, and consistent with other developmental findings, the ranges within age groups are even more striking. At every level there are children who perform like the average first-grader, one or two who perform at the level of the most talented child. Furthermore, the various skills cluster; children with good memories are also the most original and the most sensitive to style. Though it is possible that the exercise has merely tapped general intelligence or task ability, it seems more probable that a small percentage of the population is especially gifted in the verbal-literary area and that this group is already identifiable at an early age.

These two conclusions about literary skill may appear inconsistent. One the one hand, each age group can be separately characterized; on the other, a small group can be isolated as especially talented from the first. Further paradoxes are also raised by the findings. For example, the ninth graders seem less skilled than the sixth graders on a range of tasks, yet clearly the most developed skills will belong to much older individuals. Perhaps some progress toward resolving such puzzles can be made if one assumes a universal sequence to literary development which will be interrupted if the stages do not proceed at a sufficiently rapid rate. According to this view, all individuals would pass from one-line episodes to picaresque creations to some capacity at appreciating styles, handling thematic materials, and producing imaginative works. Yet if the child has reached an adolescence by the time he passes through these stages of literary development, his heightened critical faculty and self-consciousness might make him resist further explorations and fall back on formulas and "safe" approaches. Only the child who, because of superior verbal skills, intelligence, or tutelage, had passed through the sequence more rapidly would be able to achieve sufficient skill and mastery of the medium during the preadolescent years; then when he became increasingly critical, he would be less likely to find his works unacceptable and could continue his literary development. Studies of creative writers lend some support to this hypothetical trajectory of literary development (9). Whether the impressive literary skill possessed by certain first graders includes the potential for quick progress through the stages independent of external influences is a crucial question for which longitudinal studies would appear necessary. Certainly the results suggest that elementary school teachers should offer their students maximum opportunity to engage in literary creativity, since the height

of their potential may well have been reached and passed before formal instruction has ever begun.

FOOTNOTES

3. The version heard by the first graders was somewhat shortened and simplified. This procedure has no apparent effect on any of the measures except the one probing memory for details, which has accordingly been eliminated.

REFERENCES

1. Ames, L. B., "Children's Stories," Genetic Psychology Monographs, 23: 337-96, 1966.

2. Gardner, H., "Childrens Sensitivity to Painting Styles," Child Development, 41:813-21, 1970.

3. Gardner, H., "The Development of Sensitivity to Artistic Styles" Journal of Aesthetics and Art Criticism, 1971, in press.

4. Gardner, H., "From Mode to Symbol," British Journal of Aesthetics, 10:359-75, 1970.

5. Inhelder, B., Piaget, J., The Growth of Logical Thinking from Childhood to Adolescence, Basic Books, New York, 1958.

6. Lenneberg, E., The Biological Foundations of Language, Wiley, New York, 1967.

7. Machotka, P., "Le Développement des Critères Ésthetiques chez l'enfant," Enfance, 16:357-79, 1963.

8. Pitcher, E., Prelinger, E., Children Tell Stories, International Universities Press, New York, 1963.

9. Sarte, J. P., The Words, Braziller, New York, 1964.

10. White, S. H., "Evidence for a Hierarchical Arrangement of Learning Processes," in Lipsitt, L. P., Spiker, C. C. (eds.) Advances in Child Development and Behavior, Volume 2, Academic Press, New York, 1965, pp. 187-220.

GENERALIZING THE WHERRY-DOOLITTLE

BATTERY REDUCTION PROCEDURE

TO CANONICAL CORRELATION AND MANOVA

CHARLES E. HALL
Educational Testing Service
Princeton, New Jersey

ABSTRACT

The Wherry-Doolittle proceedure has been used for over 30 years to reduce the number of variables in a multiple correlation. This paper describes techniques for obtaining the same kind of reduction of number of variables in the cases of canonical correlation discriminant analysis and multivariate analysis of variance. Statistical tests comparable to those used in the Wherry-Doolittle procedure are cited.

SUPPOSE WE are given two sets of variables X and Y with the objective of predicting X from Y. The prediction is best accomplished by means of the regression equations attendant upon the canonical correlations (9). However, it is sometimes desired to reduce the number of variables in the Y set without disturbing the predictability greatly; one method of doing that reduction is the topic of this paper.

Similar problems exist in reducing the number of variables required to effect discrimination among groups or levels of treatment in a multivariate analysis of variance. A study of the discriminant problem was undertaken by Weiner and Dunn (10) who used four techniques of battery reduction including stepwise regression (but not the Wherry-Doolittle procedure) and evaluated the efficacy of the techniques through probability of misclassification.

THE WHERRY-DOOLITTLE PROCEDURE IN MULTIPLE CORRELATION

The calculation procedure for a Wherry-Doolittle battery reduction is described in Garrett's text (3). A short foray into algebra suffices to show that the technique is very simple in its logic, though complex in its calculation. The logic proceeds thusly:

1. Calculate all the correlations between the predictors and the criterion.

2. Choose the predictor with the largest criterion correlation as the most important of the predictors (say, predictor number 1).

3. Calculate the partial correlations between the remaining predictors and the criterion, removing the chosen predictor(s).

4. Choose as the next most important predictor that one of the remaining predictors which has the largest partial correlation with the criterion.

5. Compute the multiple correlation between the chosen predictors and the criterion. Determine whether the latest addition to the chosen predictors adds substantially to the previously obtained multiple correlation. If a substantial increase has been made, repeat steps 3, 4, and 5.

6. When addition of a new predictor fails to make a substantial increase in the multiple correlation, the procedure is terminated.

A statistical test is available for determining the importance of adding a given predictor to the set. As stated by Rao (7:225), this is the test of the

partial correlation between the last chosen predictor and the criterion removing the effect of the previously chosen predictors.

This procedure has been criticized by many users because the final selection of predictors is subject to considerable sampling variation. This shortcoming of the procedure can only be overcome by alternative information or shrewd guesswork by the researcher.

Several articles have been written on the vagaries of battery reduction procedures in multiple correlation. Burkett (1), Herzberg (5), and Rock and others (8) are recent examples. Herzberg discusses battery reduction in canonical correlations but manages to resolve the multiple criteria to a single criterion before invoking battery reduction processes.

ALTERNATIVE CRITERIA FOR CHOICE OF VARIABLE

It is worthwhile to note alternatives to the decision rules: to wit, step 4 of the procedure where one chooses the next variable to be added to the predictor set. The procedure calls for the inclusion of the variable with the largest partial correlation with the criterion. If we denote this partial correlation as r_{ic}, we may also choose that variable which makes the greatest reduction in the predictable variance, that reduction being $R^2 - r_{ic}^2$. Or, we may also choose that variable which has the largest F ratio in the significance tests for the correlation between variables and criterion, r_{ic}, since $F = r_{ic}^2/1 - r_{ic}^2$ x constant.

An alternative calculation routine is also available. Wherry's calculation procedure in step 4 is based on the Gauss-Doolittle method of matrix inversion which is complex and somewhat difficult. The calculations necessary to produce battery reduction are those that culminate in partial correlation. The same end can be obtained by using a modification of the square root method of factor analysis (4:102).

Suppose the matrix of correlations between the criterion (c) and the predictors (i = 12, . . ., n) are arranged in a matrix as in Table 1.

Also, suppose that predictor 1 has the largest correlation with the criterion. The effect of predictor 1 can be removed from R by the following strategem. Denote C_2 as column 2 of R containing the correlations of predictor 1 with all other variables. Form the matrix $C_2 \cdot C_2'$ and subtract it from R: R -

TABLE 1

CORRELATIONS BETWEEN PREDICTORS AND CRITERIA

$$
\begin{bmatrix}
1 & r_{ic} & r_{2c}\cdots r_{ic}\cdots r_{nc} \\
r_{ic} & 1 & r_{12} \\
r_{2c} & r_{21} & 1 & \\
\cdot & & & \cdot \\
\cdot & & & & \text{etc.} \\
\cdot & & & & \cdot \\
r_{nc} & & & & 1
\end{bmatrix}
$$

$C_2 C_2' = R*$. This matrix has 0's in column 2 and row 2. The other elements are: off diagonal $r_{ij} - r_{i1}r_{1j}$ is the ij -th entry and on diagonal the i -th entry is $1 - r_{ij}^2$. Dividing each entry by the square roots of its row and column diagonal gives its value as

$$
r_{ij}^* = \frac{r_{ij} - r_{i1}r_{1j}}{\sqrt{1 - r_{i1}^2} \cdot \sqrt{1 - r_{j1}^2}}
$$

which is the partial correlation between variables i and j adjusted for predictor variable 1.

It might also be noted that this calculation procedure does not require that the original matrix of predictor-criterion correlation, R, be of full rank as does the Wherry-Doolittle procedure.

GENERALIZATION TO CANONICAL CORRELATION

Suppose we are given two sets of variables X and Y with intercorrelations,

$$
R = \begin{pmatrix} R_{xx} & R_{xy} \\ R_{yx} & R_{yy} \end{pmatrix} \tag{c}
$$

where the within battery correlation matrices, R_{xx} and R_{yy}, are of full rank. The canonical correlations, $\mu^{\frac{1}{2}}$, between X and Y are determined from either of the determinantal equations

$$
|R_{yx}R_{xx}^{-1}R_{xy} - \mu R_{yy}| = 0 \tag{a}
$$

or

$$
|R_{xy}R_{yy}^{-1}R_{yx} - \mu R_{xx}| = 0 \tag{b}
$$

There are a few aspects of these equations that are seldom discussed but warrant review here. Using equation a we note that

(1) the trace, $tr(R_{yy})$ is the "total" variance of the variable set Y,

(2) the diagonal entries of $R_{yx}R_{xx}^{-1}R_{xy}$ are the multiple correlations of each of the Y variables on the variable set X,

(3) $tr(R_{yx}R_{xx}^{-1}R_{xy})$ is the variance due to hypothesis (or "between") and is the variance of the variable set Y that can be explained by regression from the variable set X.

In reducing the variable set Y to a smaller number of variables, it makes sense to choose that variable which is most highly related to the X set. This immediately suggests that we choose the Y variable with the highest multiple correlation with the X set. This can be determined from the diagonal of $R_{yx}R_{xx}^{-1}R_{xy}$ as mentioned above. This choice is consistent with the Wherry-Doolittle procedure since (a) it chooses the variable which makes the greatest reduction in the predictable variance, $tr(R_{yx}R_{xx}^{-1}R_{xy})$, and (b) it chooses that variable which has the largest F ratio in the tests for the (multiple) correlations, $R_{Y_iX}^2$, between the Y_i and X because $F = R_{Y_iX}^2/1 - R_{Y_iX}^2$ x constant.

With this in mind we can formulate rules for battery reduction in canonical correlation.

1. Calculate all the multiple correlations between the predictors, Y, and the criterion, X.

2. Choose that predictor with the largest multiple correlation with the criterion, X, as the most important predictor, say, Y.

3. Calculate the partialled multiple correlation between the remaining predictor and the criteria. This can be done by reducing the correlation matrix R (see equation c) as in the suggested alternative calculation procedure obtaining the matrix of partialled correlations,

$$R^* = \begin{pmatrix} R^*_{xx} & R^*_{xy} \\ R^*_{yx} & R^*_{yy} \end{pmatrix}$$

then formulating $R^*_{yx}(R^*_{xx})^{-1}R^*_{xy}$ which has the desired partialled multiple correlations on its diagonal.

4. Choose as the next most important predictor that one of the remaining predictors which has the largest partialled multiple correlation with the criteria.

5. Compute the canonical correlations between the chosen predictors and the criteria. Determine whether the latest addition to the chosen predictors adds substantially to the previously obtained canonical correlations. If a substantial increase has been made, repeat steps 3, 4, and 5.

6. When addition of a new predictor fails to make a substantial increase in the canonical correlations, the procedure is terminated.

A statistical test for the decision to terminate is available and is cited by Rao (7:467) in the form used in multivariate analysis of variance. A more complete discussion of this test follows under the section on multivariate analysis of variance.

The problems of reducing the X battery is identical to that of reducing the Y battery.

GENERALIZATION TO MANOVA AND DISCRIMINANT ANALYSIS

In multivariate analysis of variance and discriminant analysis (a one-way MANOVA), the problem of battery reduction is almost identical to the problem in canonical correlation. The logic is identical up to the slight differences in the mechanics of obtaining the solution. Once the relationship between the mechanics of the two solutions is seen, the similarity of the battery reduction procedure is apparent.

In a MANOVA or discriminant analysis the solution originates from the determinantal equation

$$|SS_H - \lambda SS_E| = 0 \qquad (d)$$

where SS_H and SS_E represent the sums of squares for hypothesis and error respectively.

For convenience sake, let us denote the variables which are measured as the Y variables.

To obtain the sums of squares for hypothesis it is often convenient to set up p dummy variables (dummy parameters, design parameters, or fixed variables) when there are p + 1 levels of the experimental design (or p + 1 groups in the discriminant analysis). Let us denote these variables as the X set. Next the sums of squares and cross products of the X variables and the Y variables are obtained as the (partitioned) matrix

$$S.P. = \begin{pmatrix} X'X & X'Y \\ Y'X & Y'Y \end{pmatrix}$$

The term SS_H is then obtained by calculating

$$SS_H = Y'X(X'X)^{-1}X'Y \qquad (e)$$

the SS_E is chosen (usually as some residual) and the determinantal equation d is solved.

Equation d can be manipulated as follows. Substituting e into d we obtain

$$|Y'X(X'Y)^{-1}X'Y - \lambda SS_E| = 0$$

Substituting $\mu/(1-\mu)$ for λ we can obtain

$$|Y'X(X'Y)^{-1}X'Y - \mu(SS_E + Y'X(X'X)^{-1}X'Y)| = 0 \quad (f)$$

The term $SS_E + Y'X(X'X)^{-1}X'Y$ is recognized as the total sum-of-squares matrix, SS_T. (Note that in a discriminant analysis this is $Y'Y$ of the S.P. matrix.) Rewrite f as

$$|Y'X(X'X)^{-1}X'Y - \mu SS_T| = 0$$

Designating σ_T as the diagonal matrix of the square roots of the diagonal elements of SS_T we can pre- and post-multiply by σ_T^{-1} and obtain

$$|\sigma_T^{-1}Y'X(X'X)^{-1}X'Y\sigma_T^{-1} - \mu\sigma_T^{-1}SS_T\sigma_T^{-1}| = 0$$

Now $\sigma_T^{-1}SS_T\sigma_T^{-1}$ is a correlation matrix, say R_{YY}, in the "total" variance of the variable Y.

If we also designated σ_X as the diagonal matrix of the square roots of the diagonal elements of $X'X$ we may rewrite the equation again as

$$|\sigma_T^{-1}Y'X\sigma_X^{-1}(\sigma_X^{-1}X'X\sigma_X^{-1})^{-1}\sigma_X^{-1}X'Y\sigma_T^{-1} - \mu\sigma_T^{-1}SS_T\sigma_T^{-1}| = 0$$

It is immediately apparent that $\sigma_X^{-1}X'X\sigma_X^{-1}$ is the correlation matrix for the X variables, say R_{XX}, and that $\sigma_T^{-1}Y'X\sigma_X^{-1}$ is a cross correlation matrix, say R_{YX}. The equation can be now written as

$$|R_{YX}R_{XX}^{-1}R_{XY} - \mu R_{YY}| = 0$$

This result shows the relationship to canonical correlations.

What remains to be shown is that the values on the diagonal of $R_{YX} R_{XX}^{-1} R_{XY}$ are "multiple correlations" which have F ratios which are univariate F ratios of each of the Y variables in the analysis.[1]

Consider a single variable y of the set Y. The hypothesis sum of squares for y is $y'X(X'X)^{-1}X'y = SS_H^y$. If SS_E^y is the error sum of squares for y, the total sum of squares for y is $SS_E^y + SS_H^y = (\sigma_T^y)^2$. Now the diagonal entry of $\sigma_y^{-1} Y'X(X'X)X'Y\sigma_y^{-1}$ is d = $1/\sigma_y y'X(X'X)^{-1}X'y 1/\sigma_y$ which reduces to

$$\frac{SS_H^y}{(\sigma_y)^2} = \frac{SS_H^y}{SS_E^y + SS_H^y}$$

If d is the multiple correlation attached to the univariate F ratio for y then $d/1 - d \cdot df(E)/df(H)$ is that F ratio.[2] Now

$$\frac{d}{1-d} \cdot \frac{df(E)}{df(H)} = \frac{SS_H^y/[SS_E^y + SS_H^y]}{1 - SS_H^y/[SS_E^y + SS_H^y]} \cdot \frac{df(E)}{df(H)} = \frac{SS_H^y}{SS_E^y} \cdot \frac{df(E)}{df(H)}$$

which is the univariate F ratio for variable y.

By this means we see that the set of rules for battery reduction in canonical correlation apply directly to discriminant analysis and MANOVA. The calculations for MANOVA battery reduction are most easily carried out when the data manipulation is carried out as if the problem were a canonical correlation, but the result is the desired one.

A STATISTICAL TEST FOR MANOVA AND CANONICAL CORRELATION BATTERY REDUCTION

Step 5 of the battery reduction procedure calls for a determination to be made as to whether a new predictor adds substantially to prediction obtained by previous selections. A statistical test for additional information from additional predictors has been developed by Rao (7:467) to be used in discriminant analysis. As has been observed in the previous section, canonical correlation and MANOVA are similar procedures and the statistical tests applicable to one apply to the other.

Without pursuing Rao's logic here it should be noted that the test reduces to an analysis of the relationship between the X variables and the chosen Y variable when the previously chosen Y variables are treated as if they were covariates. That is, in canonical correlation problems, calculate the significance test for the partialled multiple correlation between the criteria X and the chosen predictor with the previously chosen predictors as covariates; and in MANOVA, calculate the significance test for the chosen Y variable treating the previously chosen Y variables as covariates.

SHORTCOMINGS AND PRECAUTIONS

The Wherry-Doolittle procedures has been belabored often because the choice of reduced battery de-

pends upon the vagaries of the sample chosen: the repetition of the process on a new sample may yield a different reduced battery. This fault of the procedure is certain to be true when it is applied to canonical correlation and MANOVA. In the application to MANOVA, the difficulty may not prove to be as severe as in multiple correlation because the variables one is trying to predict are group membership variables (or design parameters) which are usually determinable without "error of measurement" and a MANOVA involves only sampling variation in the Y or predictor variables.

In canonical correlation it would appear that the problem is even more unstable than in multiple correlation. In the case of multiple correlations, not only are the predictors subject to sampling variation, but the one criterion variable is also. In canonical correlation, the criteria are numerous and the sampling variations among them are compounded. For this reason, it may be that battery reduction in canonical correlation is almost a useless procedure because of sampling variation. Meredith(6) has devised a procedure for correcting canonical correlation data for error of measurement in the variables and has found that the correction procedure greatly altered the results of analysis. This is a strong suggestion that battery reduction in canonical correlation may be an almost useless procedure.

COMMENTS ON EFROYMSON'S STEPWISE REDUCTION AND THE STEP-UP PROCEDURE

Efroymson's Stepwise procedure (2) differed from the Wherry-Doolittle procedure only by permitting variables to be dropped. This requires that a single step be inserted in the regimen after step 5. Before recycling steps 3, 4, and 5 to consider addition of another variable, compute the partialled correlations for each of the chosen variables holding all other chosen variables as covariates. The smallest of these partialled correlations is tested for its contribution to the criterion by the test cited by Rao. If a decision is made to drop one variable, it is eliminated and the new (reduced by one) set of chosen variables is reexamined for deletion of another variable. If no variable can be dropped, the process recycles through steps 3, 4, and 5.

The step-up procedure is the process where the entire set of predictors is examined to determine whether one or more variables can be eliminated for lack of contribution to prediction. This is the same process as for deletion of chosen variables in the Efroymson procedure if one starts with the entire set of variables as "chosen" variables. In short, consider the partialled multiple correlation between each predictor and the criteria with all other predictors as covariates. The smallest of these partialled multiple correlations is tested as cited by Rao and a decision made about dropping the variable. If a decision is made to drop the variable, the reduced set of variables is reexamined for deletion of another variable.

A FORTRAN IV subroutine is available from the author to perform the techniques discussed in the paper.

FOOTNOTES

1. This is a necessary argument because the in-

tuitive choices for battery reduction in MANOVA are based on the variables with the largest (partialled) univariate F ratios at each stage of reduction; it would be desirable to show that this is consistent with the Wherry-Doolittle procedure.

2. df(E) and df(H) denote degrees of freedom for error and hypothesis.

REFERENCES

1. Burket, G. R., "A Study of Reduced Rank Models for Multiple Prediction," Psychometric Monographs, No. 12, 1964.

2. Efroymson, M. A., "Multiple Regression Analysis," in Ralston, A.; Wilf, H. S. (eds.), Mathematical Methods for Digital Computers, Wiley, New York, 1960.

3. Garrett, H. E., Statistics in Psychology and Education, David McKay Co., New York, 1966.

4. Harman, H. H., Modern Factor Analysis, The University of Chicago Press, Chicago, Illinois, 1960.

5. Herzberg, P. A., "The Parameters of Cross-Validation," Psychometric Monographs, No. 16, 1969.

6. Meredith, W., "Canonical Correlations With Fallible Data," Psychometrika, 29:55-66, 1964.

7. Rao, C. R., Linear Statistical Inference and Its Applications, Wiley, New York, 1965.

8. Rock, D. A.; Linn, R. L.; Evans, F. R.; Patrick, C., "A Comparison of Predictor Selection Techniques Using Monte Carlo Methods," Research Bulletin RB-69-76, Educational Testing Service, Princeton, New Jersey, 1969.

9. Thompson, G., "The Maximum Correlation of Two Batteries," British Journal of Psychology, Statistical Section, 1:27-34, 1947.

10. Weiner, J. M.; Dunn, O. J., "Elimination of Variates in Linear Discrimination Problems," paper presented at the American Statistical Association meeting, Chicago, Illinois, December 27, 1964.

Book Reviews

New paradigms in education, as in other fields, create resistance. In later chapters, Borton considers some of the particular resistances to process education: parents may fear that their children will not learn academic skills; teachers and students may have long developed expectations about educational goals and methods which run counter to those necessary for involvement in a process orientation. A more subtle resistance, and one never made explicit by Borton, results from the fact that an affective education of the sort described by Borton undoubtably conjures up images of psychotherapy. Borton alludes to this potential resistance when he attempts to reassure the reader that he intends process education to be education and not therapy.

The image of therapy has a number of elements. To some it connotes pathology and mysterious therapeutic processes. Borton is presumably attempting to remove these connotations and thereby overcome resistance. It may be, however, that public resistance stems from other elements of therapy: attention to feelings and the goal of personal change. In a culture unaccustomed to expressing feelings or communicating about them, Borton's views will not be accepted easily by educators or parents. Furthermore, the goal of personal change requires skills of teachers that they have not been previously taught. Opposition to the therapy-like aspects of process education, then, may be founded on elements other than those which can be removed through simple reassurance. The problem of overcoming resistance, if this analysis is correct, is one of legitimizing the exploration of feelings and of helping educators to increase their own interpersonal skills in promoting change.

Education and therapy have always shared some common elements. The attention to feelings found in newer educational proposals and the increasing use of the definition of learning as behavioral change are adding to the similarity. Given this similarity, it might be appropriate to meet the issue head on and to explore ways in which teachers and students can be helped to live more effective, affective lives. Borton urges that schools interested in process education provide their staffs with sensitivity training. Such training allows teachers to experience, as "students," interpersonal and affective learning, and it may partially prepare teachers to understand and promote the learning of their own students. Beyond sensitivity training, there is a need for teachers to monitor their own and their students' feelings about the educational program. The systematic collection and examination of data may be the best way to ensure that teachers and students learn.

Borton's book concludes with an appendix containing reference to articles, books, and films that serve to continue the reader's own educational process. Borton's book, like his model of learning, is inductive. He begins with his own experiences, cognifies them, and applies them to an educational program. He leaves the reader with suggested materials providing for further experience and learning.

Steven R. Asher, Reviewer
Instructional Research Laboratory
University of Wisconsin-Madison

ORGANIZATIONAL CLIMATE AND FREQUENCY OF PRINCIPAL-TEACHER COMMUNICATIONS IN SELECTED OHIO ELEMENTARY SCHOOLS[1]

CARL HELWIG
Old Dominion University, Norfolk, Virginia

ABSTRACT

Often it had been stated in organizational conflict theory that when organizational homeostasis became unbalanced, the participants in conflict should "communicate more." To test this assertion, the average total frequency of principal-teacher oral and written communications over an identical 20-day period were correlated with two empirically-determined variables, namely school organizational climate, explaining the "nature" of homeostasis, and teacher esprit, the degree of teacher satisfaction or morale. Thirty-seven cooperating Ohio elementary schools and their respective faculties formed the sample. The basic assumption was not upheld ($p \leq .05$). Further refined statistical analyses of the data also failed to give additional support to the above contention. Cautious inferences were then drawn.

CENTRAL TO a changing system of interaction is the process of communication. Communication aids or hinders goal achievement within the organization and it affects group membership (3:534). Because the frequency of principal-teacher communications in the public elementary school might have been a determinant in the school's organizational climate as well as in its teacher esprit (morale), a hypothesis was tested, namely: that the total frequency of oral and written communications between the principal and his faculty collectively, as well as downward from the principal to the faculty and upward from the faculty to the principal, were significantly ($p < .05$) related to the nature of the school's organizational climate as well as the faculty's esprit.

METHODOLOGY

The nature of a school's organizational climate and the degree of its faculty's esprit can be determined through the work of Halpin. Describing the school's organizational climate as the organizational personality of the school, Halpin through factor analysis derived six profiles or prototypic organizational climates for the elementary school. These profiles, moreover, arranged themselves along a continuum from open to autonomous, controlled, familiar, paternal, and closed prototypic climates.

Three parameters were also discovered in describing the social interaction between an elementary principal and his faculty: authenticity, satisfaction, and leadership initiation. The first, said Halpin, defined the "openness" of the behavior between the principal and his faculty; the second, "the attainment of conjoint satisfaction in respect to task accomplishment and social needs"; and the third, the latitude with which the principal as well as the faculty initiated leadership acts.

In this investigation, the primary concern was with the second conceptualization, "the conjoint satisfaction in respect to task accomplishment and social needs." For the faculty, this resulted in esprit, but it was not the sole determinant in the school's organizational climate. Eight behavioral patterns, four belonging to the principal and four to the faculty, covarying among themselves, identified the school's organizational climate as being one of the 6—open, autonomous, controlled, familiar, paternal, or closed. Halpin labeled the four principal behaviors as thrust, production emphasis, aloofness, and consideration and the four faculty behaviors as esprit, intimacy, disengagement, and hindrance (4). His Organizational Climate Description Questionnaire (OCDQ) identified the school's organizational climate through these eight subdimensions.

Data on the viability of the construct, school organizational climate, were contained in the sources listed in Table 1 (5). Reliability data have been reported by Halpin and Anderson as follow in Table 1 and 2.

The acceptable reliability of the OCDQ was again demonstrated by Anderson (1:81) who in a test-retest Pearsonian r correlation, as well as an odd-even respondent Pearsonian r with a Minnesota sample, obtained the reliability coefficients ($p < .01$) shown in Table 2.

The Principal's Data Sheet (PDS) was designed to obtain the frequency of various types of oral and written communications between a principal and his faculty, but for this investigation, the average of the total frequency of communications over a 20-day period in each school within the sample became the sole measure. By type of communication in a pilot study, item reliability coefficients were significant at least at the .05 level, while the odd-even respondent reliability coefficient for the whole PDS was .82, significant at the .01 level (6:37-39).

The population consisted of the 3,107 elementary schools listed in the 1966-67 Educational Directory of the State of Ohio (7). Proportionate random sampling by type of school allowed the mailing of seventy-two requests to city schools, sixty requests to county schools, and eight requests to exempted village schools. Fifty-two principals replied that they were willing to cooperate. Thirty-seven principals actually completed the PDS, the other fifteen failing to respond to a tracer letter after the instruments had been mailed to them.

TABLE 1

HALPIN'S ESTIMATES OF INTERNAL CONSISTENCY AND OF EQUIVALENCE FOR THE EIGHT OCDQ SUBDIMENSIONS (5:49)

OCDQ Subtests	Split-half Coefficient of Reliability, Corrected by the Spearman-Brown Formula[a] (N=1,151)	Correlation Between Scores of the Odd-Numbered and the Even-Numbered Respondents in Each School[b] (N=71)	Communality Estimates[c] for Three-Factor Rotational Solution (N=1,151)
Disengagement	.73	.59	.66
Hindrance	.68	.54	.44
Esprit	.75	.61	.73
Intimacy	.60	.49	.53
Aloofness	.26	.76	.72
Production Emphasis	.55	.73	.53
Thrust	.84	.75	.68
Consideration	.59	.63	.64

[a]Estimate of internal consistency.
[b]Estimate of equivalence.
[c]These are lower-bound, conservative estimates of equivalence.

TABLE 2

ANDERSON'S RELIABILITY COEFFICIENTS

Test-Retest Pearsonian r		Pearsonian r Correlation of Odd-Even Respondents
Disengagement	+.567	+.541
Hindrance	+.458	+.791
Esprit	+.805	+.685
Intimacy	+.653	+.668
Aloofness	+.196	+.708
Production Emphasis	+.787	+.692
Thrust	+.504	+.763
Consideration	+.805	+.556

Each cooperating principal was sent ten copies of the OCDQ and asked to distribute them randomly among his faculty. The percent of return by school ranged from 70 to 100 with the exception of three schools. By thus sampling generally 50 percent or more of the eligible faculty population within each of the thirty-seven schools, a high degree of precision could be attained in inferring to the whole faculty of each school (2:3). For the total sample, 310 OCDQ's were returned of the 645 sent to the cooperating principals; this represented a 47 percent response for the total sample.

Of the thirty-seven schools in the sample, twenty-one were city schools; thirteen, county schools; and three, exempted village schools. No discernible reason could be given for the fifteen principals who failed to reply to the tracer letter other than that eight were from city schools, six from county schools, and one from an exempted village school. That these principals failed to reply may have biased the sample as well as the procedure employed, whereby each cooperating principal selected the teachers to whom he passed out the OCDQ's.

The nonparametric Spearman (rho) rank correlation coefficient was selected as the main statistic for it was a distribution free statistic and had about a 91 percent efficiency of the Pearson correlation coefficient in rejecting a null hypothesis. Since the sample, as indicated above, may have become biased, the rho correlation coefficient seemed to be the more appropriate statistic to apply to the obtained data. But, in addition, although the OCDQ itself was a summated, Likert-type, equal interval scale, the PDS, as constructed, did not meet the interval scale requirement, but involved ordinal measurement instead. Therefore, again the Spearman rho, not the Pearson r, seemed to be the more appropriate correlational statistic (9:202-213).

RESULTS

Table 3 shows the results of Spearman rank correlations between the frequency of total principal-teacher communications, the frequency of principal downward communications to the faculty, the frequency of teacher upward communications to the principal, and the OCDQ esprit mean scores. The rho correlation by school between the frequency of total principal-teacher communications and the OCDQ esprit mean scores was .21, between the frequency

TABLE 3

SPEARMAN RANK CORRELATIONS

	OCDQ Esprit Mean Scores
Frequency of Total Principal-Teacher Communications	.21*
Frequency of Principal Downward Communications to the Faculty	.28
Frequency of Teacher Upward Communications to the Principal	.31

*None of the above rho's significant at the .05 level of acceptance on one-tailed test.

of principal downward communications to the faculty, .28, and between the frequency of teacher upward communications to the principal, .31. A rho correlation of .39 was needed in each of these three instances at the .05 level of acceptance on a one-tailed test.

The sample yielded six open, five autonomous, three controlled, no familiar, five paternal, and eighteen closed climate schools.

Brown and Watkins in their own research both had raised some doubt about Halpin and Croft's intermediate school climate designations of controlled and familiar. Brown identified with his Minnesota sample all six categories of school climates, except the category, controlled climate (2:passim). Watkins with a Muscogee County School District, Georgia, sample raised some doubt about the two middle school climate categories, controlled and familiar (10:52).

The Brown and Watkins findings are mentioned in order to justify in this investigation the correlations in the open and closed school climate categories only, the extremes of the Halpin and Croft school climate continuum and not the four remaining intermediate school climate categories of autonomous, controlled, familiar, and paternal. This is also in keeping with the view of Halpin and Croft:

We have said that these climates have been ranked in respect to openness versus closed-

TABLE 4

SPEARMAN RANK CORRELATIONS BY OPEN OR CLOSED SCHOOL CLIMATE BETWEEN THE FREQUENCY OF TOTAL PRINCIPAL-TEACHER COMMUNICATIONS AND THE OCDQ ESPRIT MEAN SCORES

| Open Climate Schools | $r_S = -.09*$ | (N=6) |
| Closed Climate Schools | $r_S = .27$ | (N=18) |

*None of the above rho's significant at the .05 level of significance on a one-tailed test.

ness. But we fully recognize how crude this ranking is. As is the case in most methods of ranking or scaling, we are much more confident about the climates described at each end of this listing than we are about those described in between (5:50).

Table 4 shows the results. The rho between frequency of total principal-teacher communications and the OCDQ esprit mean scores in the open climate schools was -.09 and in the closed climate schools .27. At the .05 level of acceptance in both instances, a rho correlation of at least .83 was needed for the open climate schools and a rho correlation of at least .40 for the closed climate schools on a one-tailed test.

DISCUSSION

With no significant correlational findings (p < .05) between the total frequency of principal-teacher communications and teacher esprit, nor between the frequency of principal downward communications to his faculty and teacher esprit, nor between the frequency of teacher upward communications to their principal and teacher esprit, nor between the total frequency of principal-teacher communications and teacher esprit in the open and closed climate schools, what inferences could be safely drawn from these data?

Perhaps principal-teacher communications might involve characteristics other than merely oral or written attributes. To hold that all communication was entirely verbal communication, said Halpin, was perhaps fallacious for "actions spoke louder than words" (4:253).

Even with the low level of overt behavior herein, that is, the frequency of oral or written behavior either by the principal or his faculty, no significant differences were obtained. If overt behavior, either by the principal or his faculty, were largely communicative behavior and this in turn were related to organizational morale or climate, there was nothing in this educational setting of the elementary school with its principal and its faculty, at least with this sample, to so support such a generalization. In organizational conflict theory, some held that "people ought to communicate more" when conflict arose and thus human relations and human morale would ipso facto improve. These findings here might suggest otherwise.

If we are looking for laws of human behavior, then our concepts must be more than sets of operations, or mathematical formulas, or of logical realities, or of sheer descriptions. They must have empirical and not merely rational implications (8:268).

That during organizational conflict, the interested participants "should communicate more" might suggest a form of rationality, but this assertion must also be subjected to empirical confirmation. In this investigation, the role of conflict was not directly studied except that the closed climate school suggested little organizational homeostasis when contrasted to the open school climate. But more so, was there really a relationship between communicative behavior and any other organizational variable, including morale?

—

FOOTNOTES

1. This is a shorter version of the author's unpublished doctoral dissertation under the same title, University of Akron, 1969, and also a part of Research Grant OEG-0-8-08005-3715, "An Analysis of the Relationship of the Degree of Satisfaction of Teachers Within Certain Ohio Schools with the Formal Communication of Their Principal," Bureau of Research, Office of Education, U. S. Department Health, Education, and Welfare, Region V, Chicago, Illinois, 1969. A version of this paper was also presented under the same title at the annual meeting of the American Educational Research Association, Los Angeles, California, 1969.

REFERENCES

1. Anderson, D. P., "Relationship Between Organizational Climate of Elementary Schools and Personal Variables of Principals," unpublished doctoral dissertation, University of Minnesota, Minneapolis, 1967.

2. Brown, R. J., "Organizational Climate of Elementary Schools," Educational Research and Developmental Council of the Twin Cities of Metropolitan Area, Inc., University of Minnesota, Minneapolis, 1964.

3. Guetzkow, H., "Communications in Organizations" in Handbook of Organizations, March, J. C., (ed.), Rand McNally and Co., Chicago, Illinois, 1965.

4. Halpin, A., Theory and Research in Administration, The MacMillan Co., New York, 1966.

5. Halpin, A.; Croft, D. B., "The Organizational Climate of Schools," The Midwest Administration Center, University of Chicago, 1963.

6. Helwig, Carl, "Organizational Climate and Principal-Teacher Communications in Selected Ohio Elementary Schools," unpublished doctoral dissertation, University of Akron, Akron, Ohio, 1969.

7. Jenkins, K., (compiler), Educational Directory: State of Ohio, School Year 1966-67, Ohio State Department of Education, 1968.

8. DiRenzo, G. J., Concepts, Theory, and Explanation in the Behavioral Sciences, Random House, New York, 1966.

9. Siegel, S., Nonparametric Statistics, McGraw Hill Book Co., New York, 1956.

10. Watkins, J. F., "The OCDQ—An Application and Some Implications," Educational Administration Quarterly, 4:2, Spring 1968.

THE MEASUREMENT OF BRUNER'S

PHILOSOPHY OF CURRICULUM GOALS

PRISCILLA PITT JONES
Wheaton College
and
KENNETH J. JONES
Florence G. Heller School for Advanced Studies in Social Welfare
Brandeis University

ABSTRACT

Norms, reliabilities, and validities for the five scales of an instrument designed to measure Bruner's model are presented. In testing the instrument on a random sample of men and women, it was found that persons of both sexes emphasized knowledge of society, human condition, natural world, past, and artistic heritage in that order. It is suggested that the Educational Values Inventory may be able to provide useful information to educational planners when developing institutional goals.

IT SEEMS tenable to say that school systems traditionally have lacked effective procedures for long range planning of activities, evaluating accomplishments, and reporting results to constituents. That this situation can no longer be tolerated is evidenced by the penetrating questions concerning the goals of particular school systems, the processes of goal selection, and the degree of goal attainment being asked by students, teachers, parents, and other taxpayers. The statement of goals is of central importance in a school system because all decision making must reflect the constancy of purpose and direction made possible by such a rationale. During the last decade, the effective use of the planning-programming-budgeting system (PPBS) in various types of organizations has encouraged optimism among those concerned with educational planning (6). This optimism seems reasonable so long as new approaches to the process of goal selection are developed—for the success of PPBS or any other long range planning scheme depends heavily on the effectiveness of goal development.

The educational literature abounds with statements of goals. In general, school systems have gravitated toward two formulations. One appeared in 1946 as a result of the efforts of the Educational Policies Commission which was appointed by the National Education Association (NEA) and the American Association of School Administrators (4). The other, published in 1953, was prepared by Kearney for the Mid-Century Committee on Outcomes in Elementary Education (9). These formulations of goals are general in nature. We assume a school system attempting to develop its own goals will use such statements as guides in the development of a unique set of goals which are relevant for one particular town or city at one point in time. This is not a simple process and often the goodness of fit is not outstanding.

It is the purpose of this paper to describe an instrument which has been developed to measure the educational values of those individuals who are concerned with school systems. Goals which are uniquely appropriate for a particular school system are more likely to be developed when the values of all constituents are measured and considered. The determination of the values of those serving and those served by school systems is so vital to meaningful goal development that it should not be left to chance.

The emphasis here is on the measurement of educational values because the particular emphasis of a statement of goals appears to reflect the value orientations of the individual or group preparing them. For example, the value held by the NEA of developing the whole child so that he may fit into society is evident behind the four major categories of the Educational Policies Commission's statement on goals:

the objectives of self realization, the objectives of human relationship, the objectives of economic efficiency, and the objectives of civic responsibility (4). It is only logical that a group oriented toward social psychology would emphasize the relationships among groups of people, whereas those interested in clinical psychology would stress individual adjustment.

According to Allport, a value is a "belief upon which a man acts by preference" (1:454). Because values are relatively resistant to change, the way an individual behaves now and at some future time rests to a large degree on his personal values (1, 2). We have developed the Educational Values Inventory (8) to measure a S's educational values. Its usefulness in goal setting rests on the assumption that these values are relatively stable and are reflected in his behavior.

THE INSTRUMENT

In his essay "After John Dewey, What?" (3), Bruner presents some thoughts which are relevant to the area of goals. Based on his suggestions, it appears that it is the school's responsibility to bring about cognitive and affective changes in the student's behavior in the following areas:

1. The natural world. This area involves the student's understanding of the physical sciences and geography.

2. The human condition. Included in this domain is the student's understanding of himself—of his personality, interests, and attitudes.

3. The nature and dynamics of society. This realm consists of developing the student's awareness of and respect for such areas as (1) the feelings, opinions, and rights of individuals; (2) the values held by other people and other societies; (3) economic and political structures; and (4) civic responsibility.

4. The past. This domain involves the development of the student's understanding so that it may be used in experiencing the present and aspiring to the future.

5. The products of our artistic heritage. This dimension relates to the student's understanding and appreciation of art, music, poetry, and other creative products.

Since these areas can be understood only when the student is adept in both language and mathematics, Bruner suggests that these two tools must have a central place in the curriculum (3:121-122).

The Educational Values Inventory contains two sections. In the first part, the individual ranks from most important to least important the five areas of the curriculum noted in Bruner's model. These five areas which are described above are

1. knowledge of the natural world,
2. knowledge of the human condition,

3. knowledge of the nature and dynamics of society,
4. knowledge of the past,
5. knowledge of our artistic heritage.

In the second section, four concrete examples of each of the five areas were developed. These are

1. knowledge of the natural world:

   a. the reasons for changing weather patterns
   b. the functions of the human circulatory system
   c. magnetic fields
   d. the role of plants in the environment

2. knowledge of the human condition:

   a. why he enjoys working with his hands
   b. why he feels happy
   c. why he feels angry sometimes
   d. why he dislikes a particular person

3. knowledge of the nature and dynamics of society:

   a. the feelings of a fellow student
   b. what it is like to live in the city
   c. the values of the people in developing countries in Africa
   d. the way governmental organizations operate

4. knowledge of the past:

   a. the long term effects of Greek civilization
   b. Thomas Jefferson's role in the colonial period
   c. the reasons behind the early Scandinavian explorations
   d. the ramifications of the industrial revolution

5. knowledge of our artistic heritage:

   a. the use of rhythm in musical expression
   b. the works of Michelangelo
   c. the use of color in a painting by Van Gogh
   d. some of Robert Frost's poems.

Thus, there are four items on each of five scales. The items on the five scales were randomly rotated and paired so that an item on one scale would be presented with one item from the four remaining scales. In this way sixteen forced-choice decisions were offered for each of the five scales. This results in forty pairs of items. The following example will illustrate this procedure:

It is extremely important for the elementary school student to understand

1a. why he enjoys working with his hands
                                   scale 2, item 1
  b. the feelings of a fellow student
                                   scale 3, item 1

2a. the reasons for changing weather patterns
                                    scale 1, item 1
  b. the use of rhythm in musical expression
                                    scale 5, item 1

Although values for the elementary school were re-
quested during the administration of this instrument,
it should be generalizable to any level at least in the
Brunerian spiral curriculum model, as it is not the
topic which is a function of level but rather the depth
of investigation of that topic.

THE SAMPLE

A 10 percent random sample stratified by sex was
drawn from the March 1969 voter list of a small, up-
per middle-class community within a radius of 20
miles of Boston. The Educational Values Inventory
was mailed to 122 men and 122 women. A code num-
ber was assigned to each sample member and tele-
phone follow-ups were conducted. It was found that
the voter list was somewhat out of date and s o m e
sample members had moved from town or died. In
addition, a few were away at college, in the armed
forces, or severely ill. Out of the theoretically pos-
sible pool of Ss, fully completed questionnaires were
returned by 69 percent or forty-seven of the m a l e
sample and by 65 percent or fifty-five of the female
sample. Possible biases which would indicate that
those who responded and those who failed to respond
differ in some systematic fashion were investigated.
An analysis of the differences between respondents
and non-respondents on amount of taxes paid, num-
ber of children in school, and sex showed no statis-
tically significant differences above t h e .10 level,
suggesting no bias at least on t h e s e demographic
factors. In addition, each non-respondent was asked
to give a reason for his unwillingness t o respond.
Approximately one third declined to give a reason;
approximately one third said they would return the
questionnaire, but failed to do so; and, approximate-
ly one third were disqualified for an assortment o f
reasons. Although these results do not rule out the
possibility of bias, they make it somewhat unlikely.

ANALYSIS OF THE EDUCATIONAL VALUES IN-
VENTORY

As stated in the instrument section, the Educa-
tional Values Inventory consists of a ranking of the
general goal areas as well as a forced c h o o s i n g
among these goals in specific situations. An exam-
ination of both sections is important, as it might be
argued that although respondents may a s c r i b e to
goals in theory, they may not follow their priorities
in actual practice. Having the respondent rank the
goal areas asks him what he believes. Having him
make choices among concrete alternatives asks him
how he would behave. Naturally, a further source
of information would be to observe his behavior in
an actual situation.

The validity coefficients which indicate the cor-
relations between rank and scale are presented in
Table 1. Significant at less than the .01 level, they
suggest that the respondents make actual, behavior-
al choices consistent with their beliefs. In addition,
the results indicate that the Educational Values In-
ventory is valid with respect to Bruner's goal for-
mulation.

TABLE 1

VALIDITY COEFFICIENTS FOR EDUCATIONAL
VALUES INVENTORY (N=130)[a]

| Scale | Validity Coefficients | |
	Uncorrected	Corrected for Attenuation[b]
Natural World	.25**	.36
Human Condition	.52**	.58
Society	.32**	.55
Past	.36**	.48
Artistic Heritage	.32**	.50

[a]In addition to the scores of townspeople, scores
for groups of teachers and administrators were in-
cluded in this analysis.

[b]Assuming equal reliability of rank and scale.

**p < .01

Table 2 shows the reliability coefficients for the
five scales. These reliability coefficients w h i c h
range from .59 to .89 suggest that this instrument
has satisfactory internal reliability (5).

Due to the ipsative nature of the instrument, the
negative correlations which exist among many of the
scales are expected. These are presented in Table
3. The correlations do appear to follow a pattern
which is intuitively satisfying in that respondents in-
terested in the student's social development (Human
condition, Society) tend to be less interested in his
mastery of skills and content (Natural world, Past,
Artistic heritage).

TABLE 2

RELIABILITY COEFFICIENTS FOR THE EDUCA-
TIONAL VALUES INVENTORY[a] (N=135)[b]

Scale	Reliability Coefficient
Natural World	.68
Human Condition	.89
Society	.59
Past	.74
Artistic Heritage	.65

[a]Scale reliabilities computed from coefficient alpha,
Cronback's generalization of the Kuder-Richardson
formula 20 for continuous scales (7).

[b]See Table 1, footnote a.

TABLE 3

CORRELATION COEFFICIENTS FOR THE EDUCATIONAL VALUES INVENTORY (N=135)[a]

Scale	Correlation Coefficient				
	1	2	3	4	5
Natural World	1.00				
Human Condition	-.60**	1.00			
Society	-.33**	.15	1.00		
Past	.16	-.54**	-.35**	1.00	
Artistic Heritage	.03	-.41**	-.30**	-.07	1.00

[a]See Table 1, footnote a.

**$p < .01$

In order to facilitate the investigation of the effect of the respondent's sex on his choices, equal numbers of males and females were included in the sample. Analyses of the significance of the difference between male and female mean scores on the five scales as noted in Table 4 indicate that there is a significant ($p < .05$) sex effect only on the Artistic heritage scale. This suggests that in general male and female respondents hold the same values for students in the various goal areas. Female respondents, however, would place more emphasis on the aesthetic than males. This finding seems to be in keeping with generally accepted sex role differences.

The statistical analyses of the measurements obtained with the Educational Values Inventory do not contraindicate the satisfactory construct validity and internal reliability of this instrument.

AN APPLICATION OF THE EDUCATIONAL VALUES INVENTORY

The graphs in Figures 1 and 2 indicate that the male and female respondents value the socialization goals—Human condition and Society—more highly than the content goals—Natural world, Past, and Artistic heritage. This may be interpreted as being both fundamental and timely. Traditionally, one of

TABLE 4

SEX MEANS FOR THE SCALES OF THE EDUCATIONAL VALUES INVENTORY (Male N=47, Female N=55)

	Natural World	Human Condition	Society	Past	Artistic Heritage
Male $\overline{X}$	8.02	8.77	10.60	7.62	4.57
Female $\overline{X}$	7.40	8.73	10.27	7.56	6.04
Significance	n.s.	n.s.	n.s.	n.s.	t=2.48 $p < .02$

FIGURE 1

MEANS ON THE SCALES OF THE EDUCATIONAL VALUES INVENTORY
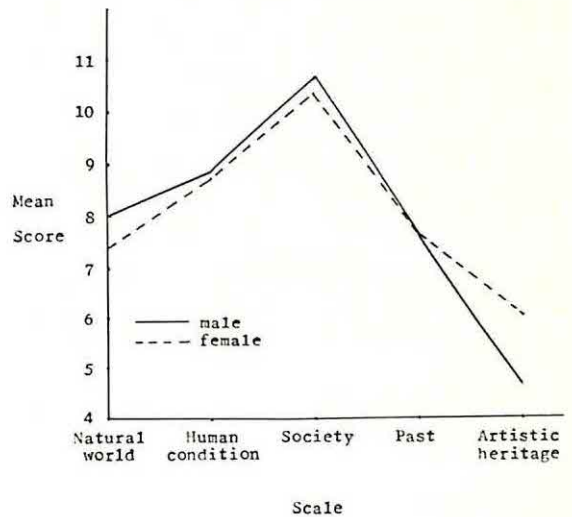
the roles of the elementary school has been socialization. Through John Dewey's work, this concept has become of fundamental importance. These results are also timely since much of today's concern centers around the stability of society and a greater concern for the realization of one's self.

In order to be consistent with the educational values of its residents, the school system of this suburban town should have a goal framework which pays

FIGURE 2

MEANS ON THE RANKS OF THE EDUCATIONAL VALUES INVENTORY
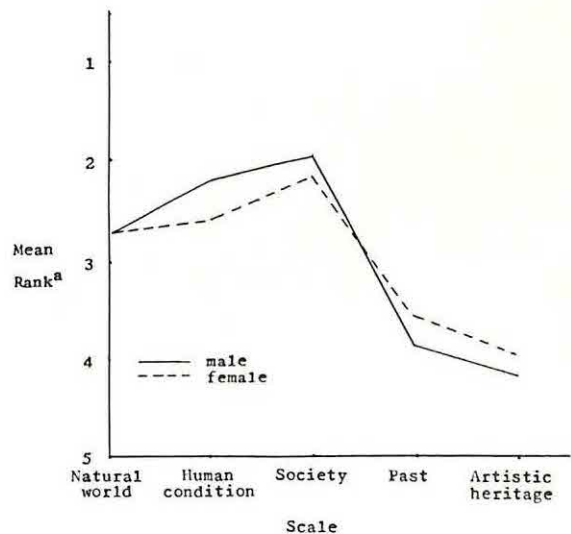
[a]The lower the value, the greater the importance.

particular attention to the socialization of its students. In addition, greater emphasis should be placed on the natural sciences and social studies than on artistic endeavors.

Naturally, the goals of a viable school system would reflect the values of the system's other constituents—students, teachers, and administrators. If the Educational Values Inventory was used to obtain data from each of these relevant groups, it would be possible to select goals on a more empirical basis than is customary. We feel that this instrument can provide useful information to the educational planner when he addresses the problem of goal formulation, something we hope will occur more frequently in the future than it has in the past.

REFERENCES

1. Allport, G. W., Pattern and Growth in Personality, Holt, Rinehart, and Winston, New York, 1961.

2. Allport, G. W.; Vernon, P. E.; Linzey, G., Study of Values, 3rd edition, Houghton-Mifflin, Boston, 1960.

3. Bruner, J., On Knowing, Harvard University Press, Cambridge, Massachusetts, 1962.

4. Educational Policies Commission, Policies for Education in American Democracy, National Education Association, Washington, D. C., 1946.

5. Guilford, J. P., Fundamental Statistics in Psychology and Education, McGraw-Hill, New York, 1965.

6. Hartley, H. J., Educational Planning—Programming-Budgeting: A Systems Approach, Prentice-Hall, Englewood Cliffs, New Jersey, 1968.

7. Jones, K. J., Multivariate Statistical Analyzer, Harvard Cooperative Society, Cambridge, Massachusetts, 1964.

8. Jones, P. P., "Educational Values Inventory," Multivariate Research Incorporated, Dover, Massachusetts, 1970 (single copies available without charge).

9. Kearney, N., Elementary School Objectives, Russell Sage Foundation, New York, 1953.

AN ALTERNATIVE

TO THE STANDARDIZED SCORE

IN GRADING A MULTIPLE-CHOICE EXAMINATION[1]

S. J. KILPATRICK, Jr.
Virginia Commonwealth University

ABSTRACT

This paper describes the current grading procedure at the Medical College of Virginia and suggests that, rather than using the standardized score to grade multiple-choice examinations, the percent of known questions be estimated and used. Standardized scores tend to be misleading when used for multiple-choice questions in that they make no allowance for guessing. It is advocated that a passing grade be awarded to those students who score significantly higher than some minimum. Under this proposal the comprehensive examination at the end of a year or phase would be replaced by reexaminations in the various subject matters but only for those students who had failed to demonstrate a sufficient grasp of this material.

THE MEDICAL College of Virginia adopted an integrated medical curriculum and a new system of grading in 1964 (3). Since then, all examinations are composed of multiple-choice questions, usually with four or five alternatives. Each student is given a pre-coded answer sheet containing his Social Security number. He records his answers by marking one of the five "boxes" against each question number. These forms are then automatically read by a computer which compares each student's answers with a master sheet, tallies the number of correct answers, and prints, in alphabetical and rank order, each student's score. This is given in four forms: the number of correct answers, the percentage of correct answers, a "z" score which is the deviation of the number correct from the class mean divided by the class standard deviation, and a standardized score which is the z score standardized to a mean of 50 and a standard deviation of 10.

Example: Consider a student, Y, who scored 66 out of 115 questions correct in a multiple-choice examination in which the class mean was 79 and the standard deviation of the class was 7.3 (Table 1). Then:

Number correct $= 66$

Percent correct $= 66/115 = 57\%$

z score $= \dfrac{66-79}{7.3} = -1.78$

Standardized score $= 50+10(-1.78) = 32$

Students are graded Honors, Pass, or Fail largely on the basis of the standardized score. While not strictly adhered to, students with standardized scores above 70 are considered for honors and those with standardized scores below 30 are considered as having failed. This is equivalent to using ± 2 standard deviations about the mean to discriminate among the three potential groups (Honors, Pass, Fail). The justification for this appears to be that about 5 percent of the normal distribution lies outside these limits. This policy is consistent with the reasoning (3) that a B grade might be awarded to those students with a standardized score between 50 and 65 (i.e., a score falling between the class mean and 1 and 1/2 standard deviations above the mean). In recent years, more use has been made of a student's rank in the examination. Since the decision to fail those with standard scores below 30 is equivalent to failing the last two or three in a class of one hundred (assuming a normal distribution of scores), there is little difference between these approaches.

The basic problem of grading under the integrated curriculum is that the committee responsible for the

examination does not (and perhaps cannot) establish the minimum passing score <u>before</u> the examination is given. One reason for this is that the material in a "subject matter" examination comes from a number of biomedical and clinical disciplines. Another reason is that all instructors in the subject matter are required to submit questions. Since the number of committee members is small compared with the number of instructors, individual members of the committee have little appreciation of the difficulty of the examination they have to grade. As a result, after the examination, the committee looks at the distribution of scores to see how students have done relative to each other.

This paper presents an approach in which the student's performance is evaluated without reference to his peers. The method attempts to estimate what a student knows about the material covered. A student would then be given a passing grade only if he had demonstrated a satisfactory mastery of the subject. The committee has to define (preferably before the examination) what is a satisfactory level of knowledge. This could be done by requiring each committee member to read those questions coming from his department and to state the minimum number he would expect a passing student to know in that section of the examination. By combining these, the committee would have arrived at a minimum number of questions a student would have to know to pass the examination. In turn, this figure could be converted into the equivalent minimum number of correct answers by substitution in equation 1.

THEORY

Assume that the student knows κ percent of the material to be examined. If the n questions in the examination are independent and are a representative sample from this material, the student may expect to know $n\kappa$ of the questions. The remaining n $(1-\kappa)$ questions he guesses, and assuming that these questions have "a" equally likely alternatives, he may expect to get $n(1-\kappa)/a$ correct by chance. His expected number correct, $E(s)$, is then

$$E(s) = n\kappa + n(1-\kappa)/a \qquad (1)$$

Rewriting the equation gives as an estimate of κ

$$\kappa = (s/n - 1/a)/(1 - 1/a) \qquad (2)$$

Now κ, the estimated knowledge, and $n\kappa$, the estimated number of questions known, are unbiased and have some advantages over the standardized score: they are more readily understood; each student's performance can be evaluated independently of his peers; confidence limits may be given for a student's knowledge of the material.

Assume that κ_0 is set at the minimum passing level of knowledge in an examination with n multiple choice questions each with a alternatives. We may calculate the probability that the ith student with the minimum passing level of knowledge κ_0 has scored s_i or greater out of n, as

$$Pr[\ s > s_i \mid n,\ a,\ \kappa_0\ ] = \sum_{s=s_i}^{n} \binom{n}{s} p^s q^{n-s}$$

where $q = (1-\kappa_0)(1-1/a)$ is the probability that a

TABLE 1

EXAMPLE OF EQUIVALENT SCORES ON AN EXAMINATION COMPOSED OF MULTIPLE-CHOICE QUESTIONS WITH FOUR ALTERNATIVES

	Class Mean	Y^b	Z^c
Number of questions	115	115	115
Number correct	79	66	64 (53 to 75)[a]
Percent correct	69	57	56 (46 to 65)
Number known	67	50 (35 to 64)	47
Percent known	58	43 (30 to 55)	41
z score	0.00	-1.78	-2.00
Standardized score	50	32	30

[a] Figures in parentheses represent 95 percent confidence limits for the estimated value.
[b] Y is a hypothetical student whose knowledge is estimated.
[c] Z represents a cutoff two s. d. below the mean (z=-2).

student knowing κ_0 of the material will answer a question incorrectly and $p = 1-q$ is its complement. Thus, in a multiple-choice examination we can tabulate along with the student's score s_i, the maximum probability that he achieve this or a higher score with an unsatisfactory level of knowledge of the subject matter. Probabilities may also be calculated that a low score is generated by a student having a passing knowledge of the subject.

APPLICATION

Consider the results obtained by a hypothetical student Y in a typical subject matter examination. The examination consisted of 115 multiple-choice questions, each with four alternatives. Y got sixty-six questions correct, giving him 57.4 percent correct. His estimated level of knowledge of this subject matter was 43.2 percent calculated from equation 2 as

$$k = (.574-.25)/.75$$

and the 95 percent confidence limits of his knowledge were (30.4%, 55.3%) calculated from equation 2 as

$$k_1 = (.478-.25)/.75 \text{ and } k_u = (.665-.25)/.75$$

where the first value in parentheses in the above formulas, viz, .478 and .665, was obtained by interpolation in Geigy Tables (2) for n=110, n=120, and a frequency of 57.4 percent. It follows that the estimated number of questions known was 115 x .432 = 50 with 95 percent confidence limits of 115 x .304 = 35 to 115 x .553 = 64. Y therefore answered sixty-six questions correctly, but of these he may have known from thirty-five to sixty-four with fifty as the most likely number of the questions known. His results are summarized in Table 1.

If a standardized score of 30 (two standard deviations below the mean) is used as the cutoff point in this examination, this is equivalent to saying that students must get more than 56 percent of the questions correct. In turn, this is equivalent to the requirement that students must be able to answer more than forty-seven questions out of the 115 without guessing or know more than 41 percent of the subject matter. To find how wide the indifference zone is, that is, for what range of scores can we not discriminate between pass and fail, we ask, what is the range of scores of 95 percent of students such as Z (Table 1) with knowledge equivalent to a standardized score of 30? These limits are calculated as fifty-three to seventy-five questions correct. The use of more than one cutoff is advocated so that, in the examination under consideration, anyone with a score below 53 would fail automatically, anyone with a score above 75 would pass automatically, and those students scoring between these limits be re-examined (1). However, this is extremely impractical under the present organization of the integrated medical curriculum. An alternative approach is to exploit the use of the comprehensive examination given at the end of the year and to break a student's score into sub-scores appropriate to the various subjects previously tested. Those students who, like Y, scored below 75 in this subject matter examination, would be informed that in the comprehensive exam-

ination they would have to demonstrate a command of that subject significantly above the minimum required. The drawback here is that, to be fair, such a comprehensive examination would involve as many questions as had been asked in each subject matter examination. The only realistic alternative appears, therefore, to replace the comprehensive examination with subject matter reexaminations taken only by those who had failed to demonstrate a satisfactory knowledge of the material earlier.

FOOTNOTE

REFERENCES

1. Cowden, D. J., "An Application of Sequential Sampling to Testing of Students" Journal of the American Statistical Association, 41:547, 1946.

2. Documenta Geigy Scientific Tables, 6th Edition, Geigy Pharmaceuticals, N. Y., pp. 99, 1968.

3. Rosinski, E. F.; Hamilton, D. L., "Examination Procedures as Part of a New Curriculum," Journal of Medical Education, 41:135, 1966.

ENVIRONMENTAL CORRELATES OF

DIVERSE MENTAL ABILITIES

KEVIN MARJORIBANKS
University of Oxford

ABSTRACT

The relationship between a refined measure of the home environment and four mental ability test scores: verbal, number, spatial, and reasoning was examined. The final sample for the study included 185 11-year-old boys and their parents. The Science Research Associates (SRA) Primary Mental Abilities test was administered to the boys. A newly constructed home interview schedule was developed and used to obtain responses from parents regarding the learning environment of the home. The environment was found to account for a large percentage of the variance in verbal and number ability and a moderate percentage of the variance in reasoning ability test scores. For spatial ability, the relationship with the environment was less definite. It was also found that the environment accounted for more of the variance in the mental ability scores than did a set of social status indicators and family structure variables.

MUCH OF THE research that has investigated the relationship between the environmental background of children and intellectual ability has concentrated on using global indicators of the environment and intellectual ability. When the environment has been defined in terms of social status characteristics, such as the occupation of the father and the education of parents or family structure variables such as the family size and crowding ratio of the home, only a relatively small proportion of the variability in the intellectual performance of children has been explained. Also, the utilization of global intelligence test scores obscures many important differences among children.

Therefore the purpose of this present study was to examine the relationship between a refined measure of the home environment and a set of mental ability test scores.

METHOD

Mental Abilities

In the study four mental abilities were examined: verbal, number, spatial, and reasoning. The mental abilities were operationalized by the scores on the relevant SRA Primary Mental Abilities subtests (1962 Revised Edition).

Environment

The environment was defined as being composed of a complex network of forces which surround the individual. It is assumed that a subset of the total network of environmental forces is related to each human characteristic. Thus for verbal, number, spatial, and reasoning ability it is proposed that sub-environments or subsets of environmental forces which will be related to each of the mental abilities can be identified. The union of the four sub-environments, which were postulated to be related to the four mental abilities, was defined as the learning environment. This learning environment may be present in the home, school, and community. Of these, the home produces the first and perhaps the most powerful influence on the development of the mental abilities. As a result, the home was chosen as the focus of the present study.

From a review of relevant theoretical and empirical literature(1, 2, 3, 4, 5, 6, 7), a set of eight environmental forces were identified. Subsets of these forces were postulated to be related to the mental abilities. These forces were labeled:

1. press for achievement

2. press for activeness

3. press for intellectuality

4. press for independence

5. press for English

6. press for ethlanguage

7. mother dominance

8. father dominance

Each of the environmental forces was defined in terms of a set of environmental characteristics which were assumed to be the behavioral manifestations of the environmental forces. A list of the environmental forces and the environmental characteristics is presented in Table 1.

The environmental characteristics that are listed in Table 1 facilitated the development of an instrument for the study. The instrument, which was in the form of a semi-structured home interview schedule, was used to gain a measure of the learning environment of the home. Thus, the environmental forces were operationalized as the scores on the relevant environmental measures constructed for the study.

The Sample

Approximately five hundred 11-year-old boys were tested, using first the California Test of Mental Maturity (CTMM) and then the SRA Primary Mental Abilities Test. The first test-taking situation was used to establish examiner-examinee rapport, to insure that all students were able to understand the test instructions, and to establish as far as possible uniform test-taking situations. The boys were assigned to two categories, one classified as middle class and the other as low class. The social class classification was based on an equally weighted combination of the occupation of the head of the household and a rating of his (or her) education. As far as possible two parallel pools of boys were formed. The purpose of the substitute pool was to provide a set of alternate families which could be used in the study if families from the first pool did not agree to participate.

The final sample consisted of ninety boys and their parents classified as middle class and ninety-five classified as low class.

HYPOTHESES

In the development of the study it was postulated that subsets of environmental forces which would be related to the mental abilities could be identified. Therefore the following hypothesis was investigated:

Hypothesis 1: The verbal, number, spatial, and reasoning ability test scores will be significantly related to subsets of scores of environmental forces.

It was also proposed that the utilization of subsets of environmental forces was a means of moving beyond the use of gross classificatory variables such as social status indicators and family structure characteristics as measures of the environment. The advantage of using the sub-environment approach was investigated by examining the following hypothesis.

Hypothesis 2: Scores on the environmental forces will be more highly related to measures of verbal, number, spatial, and reasoning ability than will other environmental measures such as social status indicators and family structure variables.

TABLE 1

THE ENVIRONMENTAL FORCES AND THEIR RELATED ENVIRONMENTAL CHARACTERISTICS USED IN THE INTERVIEW SCHEDULE

Environmental Force	Environmental Characteristics
1. Press for Achievement	1a. Parental expectations for the education of the child 1b. Social press 1c. Parents' own aspirations 1d. Preparation and planning for child's education 1e. Knowledge of child's educational progress 1f. Valuing educational accomplishments 1g. Parental interest in school
2. Press for Activeness	2a. Extent and content of indoor activities 2b. Extent and content of outdoor activities 2c. Extent and purpose of the use of T.V. and other media
3. Press for Intellectuality	3a. Number of thought provoking activities engaged in by children 3b. Opportunities made available for thought provoking discussions and thinking 3c. Use of books, periodicals, and other literature
4. Press for Independence	4a. Freedom and encouragement to explore the environment 4b. Stress on early independence
5. Press for English	5a. Language usage and reinforcement 5b. Opportunities available for language (English) usage
6. Father Dominance	6a. Father's involvement in child's activities 6b. Father's role in family decision making
7. Mother Dominance	7a. Mother's involvement in child's activities 7b. Mother's role in family decision making
8. Press for Ethlanguage	8a. Ethlanguage usage and reinforcement 8b. Opportunities available for ethlanguage usage

RESULTS

Before examining the hypotheses of the study it was considered desirable to investigate the reliability of the home environment schedule constructed for the study.

The reliability coefficients for each scale are shown in Table 2. The coefficients were estimated by determining coefficient alpha (7).

Since the study is concerned with the size of the correlations between environmental forces and mental abilities, it was considered that the reliability coefficients were of an acceptable level.

Hypothesis 1: The verbal, number, spatial, and reasoning ability test scores will be significantly related to subsets of scores of environmental forces.

The first analysis of the hypothesis involved an examination of the zero-order correlations between the scores of the four mental ability tests and the scores of the environmental forces. These latter scores were computed from a simple summation of the scores of the environmental characteristics which were used to define the environmental forces. The zero-order correlations are presented in Table 3.

The results in Table 3 indicate that the parental dominance dimensions had either low or negligible relationships with the mental abilities. To investigate the relationship between the two parental dimensions and the other environmental forces, a principal component analysis of the eight environmental forces was conducted. The unrotated factor loading matrix of the interrelations among the forces is presented in Table 4. Only those factors with an eigenvalue greater than unity have been included. The third factor had an eigenvalue of .65.

It can be observed from Table 4 that all of the environmental forces load strongly on the first factor. This general factor was labeled the learning environment of the home factor. The second factor, which loads heavily on the parental dominence forces, was labeled the parental dominance factor.

When interrelationships between the scores on the two factors and the mental ability test scores were examined it was found that the scores on the learning environment of the home factor were significantly related to scores on each of the mental abilities. None of the relationships between the scores on the parental dominance factor and the scores on the mental ability tests reached statistical significance.

TABLE 2

RELIABILITY COEFFICIENTS OF THE ENVIRONMENTAL SCALES (N = 185)

	Reliability Coefficient	Number of Items	Standard Deviation of Scores
Press for Achievement	.94	50	35.18
Press for Intellectuality	.98	18	17.05
Press for Activeness	.80	25	11.29
Press for Independence	.71	16	8.72
Press for English	.93	20	17.83
Press for Ethlanguage	.90	15	14.4
Father Dominance	.67	22	9.22
Mother Dominance	.66	22	10.33

TABLE 3

INTERRELATIONSHIPS BETWEEN THE MENTAL ABILITY TEST SCORES AND THE SCORES OF THE ENVIRONMENTAL FORCES (N = 185)

Environmental Force	Abilities			
	Verbal	Number	Spatial	Reasoning
Press for Achievement	.66**	.66**	.28**	.39**
Press for Activeness	.52**	.41**	.22**	.26**
Press for Intellectuality	.61**	.53**	.26**	.31**
Press for Independence	.42**	.34**	.10	.23**
Press for English	.50**	.27**	.18**	.28**
Press for Ethlanguage[1]	.35**	.24**	.09	.04
Father Dominance	.16*	.10	.09	.11
Mother Dominance	.21**	.16*	.04	.04

*$p < .05$
**$p < .01$

1. Ethlanguage refers to any language other than English used in the home.

Because of: (1) the exploratory nature of the study in identifying sub-environments for mental abilities, and (2) the presence of a general factor, it was decided to utilize the eight environmental forces as the sub-environment for each mental ability.

The relationship between the learning environment of the home and each mental ability was examined by computing the multiple correlation between the eight environmental forces and each mental ability. In this analysis the environmental forces formed a predictor set and the mental abilities formed the criterion vectors. The results of this analysis are presented in Table 5.

The results in Table 5 indicate that when the environmental forces are combined into a set of predictors they account for a large percentage of the variance in verbal and number ability test scores and a moderate percentage of the variance in the reasoning ability test scores. For spatial ability, the corrected multiple correlation did not reach statistical significance.

Thus the analysis of the data supports the hypothesis that verbal, number, and reasoning abilities are related to subsets of environmental forces. For spatial ability, the relationship with the environment, as measured in this study, was less definite.

TABLE 4

UNROTATED FACTOR LOADING MATRIX OF THE
ENVIRONMENTAL FORCES

Environmental Force	Factors		
	1	11	h²
Press for Achievement	.83	.06	.69
Press for Activeness	.90	.08	.82
Press for Intellectuality	.91	.01	.83
Press for Independence	.64	-.25	.47
Press for English	.76	-.10	.58
Press for Ethlanguage	.75	.10	.57
Mother Dominance	.40	.84	.87
Father Dominance	.41	-.84	.88
Eigenvalues	4.196	1.533	
Percentage of Variance Account for	52.4	19.2	
Cumulative Percentage of Total Variance	52.4	71.6	

Hypothesis 2: Scores on the environmental forces
will be more highly related to measures of ver-
bal, number, spatial, and reasoning ability than
will other environmental measures such as social
status indicators and family structure variables.

In Table 6 the zero-order interrelationships be-
tween a set of gross classificatory measures of the

TABLE 5

MULTIPLE CORRELATIONS OF EACH OF THE
MENTAL ABILITY SCORES WITH THE EIGHT
ENVIRONMENT FORCES

Mental Ability	Multiple Correlation R	Corrected[a] Multiple Correlation Rc	Percentage of Total Variance Rc²
Verbal	.72***	.71***	50.4***
Number	.72***	.71***	50.4***
Spatial	.32**	.26	6.7
Reasoning	.43***	.40***	16.0***

***p < .001
**p < .01
*p < .05
[a]Corrected to allow for cumulative errors in multi-
ple R, and for small sample size.

environment and each of the mental abilities have
been presented.

A qualitative inspection of Tables 3 and 6 indi-
cates that, in general, the environmental force
scores are more highly related to the mental ability
test scores than are the gross indicators of the en-
vironment.

A set of multiple correlation analyses was con-
ducted in order to compare the effectiveness of the
environmental force scores and the gross indicators
as predictors of the mental ability test scores. In
these analyses the amount of variance that could be
attributed to the environmental forces was computed
after accounting for the variance that could be attrib-
uted to the gross indicators of the environment. The
results of the analyses are presented in Table 7.

The results in Table 7 indicate that the learning
environmental forces account for 25 percent of the
variance in verbal ability test scores, 34 percent
of the variance in number ability test scores, and
12 percent of the variance in reasoning ability test
scores after the variance due to the combination of
social status characteristics (occupation of father,
education of father, education of mother) and family
structure variables (number of children, ordinal po-
sition, crowding ratio) has been allowed for. For
the spatial ability test scores the corrected multiple
correlation coefficient for "environment" did not
reach statistical significance.

Thus, the results provide support for the general
acceptance of the second hypothesis.

CONCLUSION

The results indicate the efficacy of utilizing the

TABLE 6

INTERRELATIONSHIPS BETWEEN GROSS INDI-
CATORS OF THE ENVIRONMENT AND MENTAL
ABILITY TEST SCORES (N = 185)

Gross Indicators	Mental Abilities			
	Verbal	Number	Spatial	Reasoning
Education of Father	.29**	.27**	.26**	.22**
Education of Mother	.39**	.33**	.21**	.16*
Occupation of Father	.43**	.30**	.31**	.29**
Number of Children in family	-.32**	-.33**	-.04	-.03
Crowding Ratio	-.34**	-.34**	-.07	-.09
Ordinal Position in family	-.26**	-.25**	-.04	-.04

**p < .01
*p < .05

TABLE 7

RELATIONSHIP BETWEEN MENTAL ABILITIES, ENVIRONMENTAL FORCES, AND GROSS INDICATORS OF THE ENVIRONMENT

Criterion	Predictor Variables	Computed Multiple Correlation R	Corrected Multiple Correlation Rc	Percentage of Total Variance Rc^2
Verbal Ability	A=6 status variables+8 environmental forces	.74***	.71***	51.0***
	B=6 status variables	.53***	.51***	26.0***
	C=A-B			25.0***
	=environment			
Number Ability	A=6 status variables + 8 environmental forces	.72***	.71***	50.0***
	B=6 status variables	.42***	.40***	16.0***
	C=A-B			34.0***
	=environment			
Spatial Ability	A=6 status variables + 8 environmental forces	.38**	.36*	13.0*
	B=6 status variables	.31***	.28***	8.0*
	C=A-B			5.0
Reasoning Ability	A=6 status variables + 8 environmental forces	.47***	.42**	18.0**
	B=6 status variables	.29**	.25	6.0
	C=A-B			12.0**
	= environment			

***$p < .001$
**$p < .01$
*$p < .05$

sub-environment approach in analyzing the relationship between the environment and intellectual performance.

The study has theoretical, methodological, and practical significance. Evidence concerning the environmental correlates of diverse mental abilities is central to theory construction in developmental psychology. Such evidence provides a clarification of both the basic nature and function of the mental abilities themselves and the characteristics of the environmental conditions that influence their development.

Methodologically the study has significance as a new instrument was developed in order to assess environmental variation. The results also relate to the practical efforts that are being made to determine the optimal educational conditions for children from diverse environmental backgrounds. If schools are to complement the home background of students it is necessary to know the specific effects that student background factors have on intellectual functioning.

Thus the study in its investigations of the relationship between the environment and mental abilities has theoretical, methodological, and practical significance. The results also indicate that it is possible to move beyond the use of global indicators of the environment to a much more detailed assessment of environments.

REFERENCES

1. Bloom, Benjamin S., Stability and Change in Human Characteristics, John Wiley and Sons, Inc., New York, 1964.

2. Coleman, James S., and others, Equality of Educational Opportunity, U.S. Department of Health, Education, and Welfare, Office of Education, No. FS5.238:38001. Superintendent of Documents, Government Printing Office, Washington D.C., 1966.

3. Dave, Ravindrakumar H., "The Identification and Measurement of Environmental Process Variables that are Related to Educational Achievement," unpublished PhD dissertation, University of Chicago, Chicago, Illinois, 1963.

4. Plowden, Bridget, and others, Children and their Primary Schools, A Report of the Central Advisory Council for Education, Her Majesty's Stationery Office, London, England, 1967.

5. Weiss, Joel, "The Identification and Measurement of Home Environmental Factors Related to Achievement Motivation and Self Esteem," unpublished PhD dissertation, University of Chicago, Chicago, Illinois, 1969.

6. Wolf, Richard M., "The Identification and Measurement of Environmental Process Variables Related to Intelligence," unpublished PhD dissertation, University of Chicago, Chicago, Illinois, 1964.

A SEMANTIC DIFFERENTIAL INSTRUMENT

FOR MEASURING ATTITUDE

TOWARD MATHEMATICS

EARL L. McCALLON
North Texas State University

JOHN D. BROWN
Southern Methodist University

ABSTRACT

A semantic differential was developed to measure attitude toward mathematics. This instrument was contrasted with a Likert type attitude instrument constructed by Aiken and Dreger (1, 2). There were 68 male and female subjects, all of whom were non-mathematics majors enrolled in a required doctoral level statistics class in the College of Education of a large state university in the southwest. The correlation between the two instruments was r = .90. It was concluded that the semantic differential constructed for this study was as effective a measure of attitude toward mathematics as the Likert type instrument. Analysis of the data also indicated that people possessing favorable and unfavorable attitudes toward mathematics differed to the greatest extent on the evaluative scales of the semantic differential, thus lending construct validity to the semantic differential.

AS INDICATED by Shaw and Wright (5), attitudes are the end products of the socialization process, and significantly influence man's responses to cess, and significantly influence man's responses to cultural products, to other persons, and to groups of persons. If one can determine the attitude of a person toward a given object, or class of objects, then this information can be used in conjunction with situational and other dispositional variables to predict and explain reactions of the person to that class of objects. To the extent that principles governing the change of attitudes can be known, they may be used to manipulate an individual's reactions to relevant objects. It is not at all surprising, then, that the study of attitudes has occupied a central place in the study of attitudes, psychology, psychotherapy, and social education, psychology, psychotherapy, and social psychology during the past 50 years.

Two widely used approaches to the measurement of attitudes are the Likert technique and the semantic differential technique. Researchers and program evaluators in education have tended to emphasize the Likert technique (3), an older but not necessarily more established approach.

PURPOSE

The purpose of this study was to contrast two attitudinal instruments for measuring attitudes toward mathematics. These instruments consisted of a Likert-type attitude instrument constructed by Aiken and Dreger (1, 2) and a semantic differential constructed by the authors. It was hypothesized that the more easily constructed semantic differential could be used to measure attitude toward mathematics as effectively as the Mathematics Attitude Scale (MAS) developed using the more involved and difficult Likert technique.

A second purpose was to determine whether the evaluative scales on the semantic differential, as determined by factor analysis, represented those scales on which the greatest scale mean differences occurred between people who view mathematics favorably and those who view mathematics unfavorably.

METHODOLOGY

Sample and Instrumentation

The sample consisted of sixty-eight male and female graduate students at a large state university in the southwest. All were students who were non-mathematics majors and were taking a required statistics course in the College of Education.

The present study made use of the MAS developed by Aiken and Dreger (1, 2). This instrument is an opinionnaire which makes use of a 5-point scale ranging from strongly disagree to strongly agree on each item. The MAS consists of twenty items with ten items stated positively and ten items stated negatively.

Some examples of the positive items are Item 4, "Mathematics is fascinating and fun," and Item 11, "Mathematics is something which I enjoy a great deal." An example from the negative items is Item 7, "I feel a sense of insecurity when attempting mathematics."

The instrument is scored to reflect a positive attitude toward mathematics by assigning a 1 to strongly disagree and a 5 to strongly agree on the positive items and conversely on the negative items. A score is obtained by summing the value assigned to a S's response on each of the twenty items.

The test-retest reliability for the instrument, according to Aiken and Dreger (1, 2) is r = .94. Content validity is assumed; however, a test of independence between the scores on the attitude scale and scores on four items designed to measure attitudes toward academic subjects, in general, suggested that attitudes specific to mathematics were being measured. Aiken and Dreger also established predictive validity coefficients (mathematics achievement) of .67 and .63 for males and females respectively. Both coefficients were found to be statistically significant.

Semantic Differential Scales of the type developed by Osgood (4) have proven useful to researchers in quantifying highly subjective data. The semantic differential used in this study was designed to measure attitude toward mathematics. The concept used was MATHEMATICS. The instrument consisted of fifteen bipolar adjectives placed at opposite ends of a 7-point continuum, e. g.:

Pleasant__:__:__:__:__:__:__Unpleasant

The adjectives used were: Pleasant - Unpleasant; Bad - Good; Hard - Soft; Afraid - Unafraid; Active - Passive; Valuable - Worthless; Strong - Weak; Love - Hate; Fast - Slow; Comfortable - Uncomfortable; Awful - Nice; Enjoyable - Unenjoyable; Light - Heavy; Varied - Repetitive; and Secure - Insecure.

Upon analysis of the scales, higher scores, or more favorable scores, resulted directly from the extent to which the perceived entity was rated closest to the following poles: Pleasant, Good, Soft, Unafraid, Active, Valuable, Strong, Love, Fast, Comfortable, Nice, Enjoyable, Light, Varied, and Secure.

The semantic differential was constructed according to the criteria given by Osgood (4). Unlike the construction of the MAS, elaborate item analysis procedures and repeated revisions of the semantic differential instrument were not necessary. This constitutes a major advantage of semantic differential technique.

TABLE 1

MEANS AND STANDARD DEVIATIONS FOR SEMANTIC DIFFERENTIAL SCALES

Bipolar Adjectives	Mean	Standard Deviation
Pleasant-Unpleasant	4.4706	1.9276
Bad-Good	4.9559	1.7142
Hard-Soft	2.8235	1.3376
Afraid-Unafraid	4.2059	1.9122
Active-Passive	4.8676	1.7611
Valuable-Worthless	5.8824	1.3442
Strong-Weak	4.8088	1.7557
Love-Hate	4.2206	1.4439
Fast-Slow	4.2059	1.7239
Comfortable-Uncomfortable	4.1618	1.7923
Awful-Nice	4.4118	1.5572
Enjoyable-Unenjoyable	4.5000	1.7577
Light-Heavy	3.1765	1.5056
Varied-Repetitive	4.8382	1.7157
Secure-Insecure	4.0000	1.7364

Procedure

The data were collected during the summer of 1970. The sixty-eight male and female Ss, all of whom were non-mathematics majors, were enrolled in a required doctoral level statistics class in the College of Education. One-half of the Ss were given the MAS. Approximately one week later, these Ss were given the semantic differential designed by the authors. The same procedure was followed with the other half of the group except the order of the instruments was reversed.

Treatment of Data

Means and standard deviation for all fifteen scales on the semantic differential are presented in Table 1. To determine the evaluation scales on the semantic differential, a factor analysis was performed. The results of this analysis appear in Table 2. Two factors emerged. Factor I appeared to be the evaluative factor and was associated with 57.4 percent of the explained variance. Factor II, a potency factor, accounted for 10.6 percent of the explained variance. The bipolar adjective pair, secure-insecure, appeared to be associated with both factors.

A summative rating was obtained for each subject on the MAS, the eleven evaluative scales on the

TABLE 2

FACTORS AND FACTOR LOADINGS RESULTING FROM SUBJECTS EVALUATION OF MATHEMATICS[a]

Bipolar Adjectives	Factor I	Factor II	h^2
Pleasant-Unpleasant	.72	.53	.80
Good-Bad	.85	.27	.81
Active-Passive	.81	.15	.68
Valuable-Worthless	.75	.12	.57
Strong-Weak	.85	.10	.73
Love-Hate	.81	.32	.76
Fast-Slow	.62	.17	.41
Comfortable-Uncomfortable	.71	.49	.74
Awful-Nice	.75	.47	.78
Enjoyable-Unenjoyable	.73	.50	.78
Varied-Repetitive	.56	.04	.31
Hard-Soft	.14	.79	.64
Afraid-Unafraid	.43	.66	.63
Light-Heavy	.08	.82	.67
Secure-Insecure	.64	.64	.83

[a]Varimax rotation

semantic differential, and a total semantic differential score, composed of all fifteen scales. The intercorrelations of these measures are presented in Table 3.

Although a factor analysis was performed to identify the evaluative scales, validity consideration (of the semantic differential) dictated another approach to the analyses of the data.

TABLE 3

INTERCORRELATIONS AMONG THE MAS AND SEMANTIC DIFFERENTIAL SCORES

Measures	Intercorrelations[a]		
MAS	1.00	.87	.90
SD (Evaluative Scales)		1.00	.97
SD (All Scales)			1.00

[a]All correlations are significant at the .001 level

TABLE 4

DIFFERENCES BETWEEN MEAN RATINGS ON SEMANTIC DIFFERENTIAL SCALES FOR FAVORABLE AND UNFAVORABLE ATTITUDE GROUPS

Scale	Group A[b] Mean	Group B Mean	Difference
Pleasant-Unpleasant (E)[a]	5.58	2.56	3.02
Enjoyable-Unenjoyable (E)	5.51	2.76	2.75
Comfortable-Uncomfortable (E)	5.16	2.44	2.72
Good-Bad (E)	5.86	3.40	2.46
Awful-Nice (E)	5.25	2.96	2.29
Active-Passive (E)	5.69	3.44	2.25
Strong-Weak (E)	5.60	3.44	2.16
Afraid-Unafraid (P)	4.93	2.96	1.97
Love-Hate (E)	4.93	3.00	1.93
Fast-Slow (E)	4.88	3.04	1.84
Valuable-Worthless (E)	6.37	5.04	1.33
Hard-Soft (P)	3.20	2.16	1.04
Varied-Repetitive (E)	5.16	4.28	.88
Light-Heavy (P)	3.44	2.72	.72

[a](E) represents an evaluative scale, (P) a potency scale.
[b]Group A had a favorable attitude toward mathematics. Group B had an unfavorable attitude toward mathematics.

The Ss in the study were divided into two groups on the basis of the MAS: those with a favorable attitude toward mathematics and those with an unfavorable attitude. Means on the fifteen scales of the semantic differential were computed for each group. Table 4 presents the bipolar adjectives for each of the two factors arranged from largest mean-scale differences to smallest. The bipolar adjective pair, secure-insecure, was omitted since it did not belong to a single factor.

CONCLUSION

As the data presented in Table 3 indicate, there is a high positive correlation (r = .90) between the total score on the semantic differential and the score on the MAS. There is also a high positive correlation (r = .87) between the total score on the

evaluative scales of the semantic differential and the score on the MAS. These correlations are significant at the .001 level.

It was therefore concluded that the semantic differential constructed for this study proved to be as effective a measure of attitude toward mathematics as the MAS. Considering the ease with which the semantic differential was constructed and the fact that no extensive refinement of the instrument was necessary, application of the semantic differential technique would appear to be a more practical approach to the measurement of attitudes in mathematics.

From the data presented in Table 4, it is evident that the evaluative scales as determined by factor analysis did indeed represent those scales on which the greatest scale mean differences occurred between people who viewed mathematics favorably and those who view mathematics unfavorably. Thus, construct validity for the semantic differential can be inferred in that it exhibited both internal (factor analysis) and external (correlational) validity.

SUMMARY

A semantic differential constructed by the authors was contrasted with a Likert attitude instrument, the MAS, constructed by Aiken and Dreger (1, 2). The sample consisted of sixty-eight graduate students who were not mathematics majors, but were required to take a statistics course as a part of their graduate study.

Correlation coefficients computed among the sets of total semantic differential scores, the semantic differential evaluative scales scores, and the MAS scores indicated the semantic differential was as effective a measure of attitude toward mathematics as the MAS. The significance of these relationships is discussed.

Further analysis of data substantiated the hypothesis that people possessing favorable and unfavorable attitudes toward mathematics would differ to the greatest extent on the evaluative scales of the semantic differential, thus lending construct validity to the semantic differential.

REFERENCES

1. Aiken, L. R.; Dreger, R. M., "The Identification of Number Anxiety in College Populations," Journal of Educational Research, 68:344-351, December 1957.

2. Aiken, L. R.; Dreger, R. M., "The Effects of Attitude on Performance in Mathematics," Journal of Educational Psychology, 52:19-24, 1961.

3. Likert, Rensis, "A Technique for the Measurement of Attitudes," Archives of Psychology, No. 140, 1932.

4. Osgood, C. E.; Suci, G. J.; Tannenbaum, P. H., The Measurement of Meaning, University of Illinois Press, Urbana, 1957.

5. Shaw, M. E.; Wright, J. M., Scales For the Measurement of Attitudes, McGraw-Hill, New York, 1967.

BIRTH ORDER, INCOME, SEX, AND

SCHOOL RELATED ATTITUDES

ROBERT F. McCLURE
University of Kentucky

ABSTRACT

Birth order was used to predict grades and school related attitudes following Bradley's hypothesis that first-borns are more academically interested and less socially interested than later borns. Income, sex of child, and size of family were studied as interacting variables. Several interaction effects were found in the absence of birth order main effects which indicate a need to study birth order as an interacting variable rather than as a single independent variable.

A NUMBER OF studies have related academic behavior to a child's ordinal position in the family. Schacter (6) and Sampson (5) have shown that first born children are overrepresented in college populations, have higher grade point averages than (GPA) and higher need achievement scores than later borns. Bradley and Sanborn (2) examined high school teacher selections for "superior student" counseling and found that first borns were significantly overrepresented. They concluded that social behaviors and attitudes of the children influenced the teacher selections, since there were no birth order intelligence differences.

Bradley (1), in a review, cited evidence that first borns were more adult oriented, more serious, and less popular than later borns. He concluded that first borns were more exposed to adult values and social pressures because of their ordinal position, hence, would reflect these pressures more than later borns. Bradley also hypothesized that first borns would perform better academically, be more interested in academic activities, and be less interested in extracurricular activities than later borns because of the same social pressures.

Campbell (3) and McClure (4) have stated that first borns are overrepresented in college populations, but have given evidence that GPA and school related behaviors and attitudes are not related in a simple way to birth order.

The present study used Bradley's hypotheses about birth order effects to predict grades and academic attitudes. It further studied the interaction effect of sex of child, income of father, and size of family with birth order in an attempt to find out what variables cause differences in school related attitudes.

METHOD

Subjects were 372 freshman and sophomore psychology students of both sexes. An anonymous questionnaire with fourteen questions was answered within 10 minutes at the start of a regular class period. The first six questions established background data such as age, sex, birth order, size of family, and GPA. The next four questions used a 10-point rating scale to assess attitudes about academic and extracurricular activities, which are described in McClure (4). The last four questions gave Ss an opportunity to select one of three behavior preferences such as a choice among "living alone, living with a roommate, or living in a fraternity." The only instructions given for the questionnaire were: "Please fill out this questionnaire anonymously for some research that is being done."

RESULTS

There were no main effects of birth order to predict grades or attitudes, although there were several trends in the predicted directions. For example, first borns had higher mean GPA's than later borns by chi-square analysis[1] (X^2=16.49, df=9, p=.10). Also, first borns reported a trend toward more academic enjoyment than later borns (X^2=15.02, df=9, p=.10).

There were several interesting interaction differences which were significant, however. College GPA's of first and later born males were compared by chi-square, with every S having completed at least one semester of college. This revealed that first born males had a significantly higher average than later born males ($X^2=18.77$, df=9, p=.05). However, when the mean GPA's (2.65 and 2.48, respectively) were compared by analysis of variance only a trend at the .10 level was found (F=3.57, df=1/176). There was no significant difference in GPA for first and later born girls ($X^2=5.96$, df=9, nonsignificant).

The academic attitude scales were all nonsignificant, but the behavior preference questions showed definite differences between first and later borns.

When upper income ($11,000 plus) children were given a choice among reading a book, watching TV with a friend, or talking to a friend, first borns were significantly more likely to read a book and later borns to talk to a friend ($X^2=6.25$, df=2, p=.05). This difference did not occur with lower income children.

When large family (4 or more siblings) children were given a choice among the same behaviors, first borns chose to read a book and later borns to watch TV with a friend or talk to a friend ($X^2=6.73$, df=2, p=.05). This difference did not occur in smaller family sizes.

When large family children were given a choice among watching a movie, going to a football game, or going to a party, first borns showed a trend to chose a movie, while later borns chose the party ($X^2=5.91$, df=2, p=.06). There were other similar trends, which were in the direction of Bradley's hypotheses, but nonsignificant.

These findings indicate several things. First, there may be attitudinal and behavior choice consequences of being first born which influence school achievement. Second, these attitudes may not be related in any simple way to birth order. There were interactions with sex, income, and size of family in this study before birth order effects became significant. Third, although the results were in the direction hypothesized by Bradley and others, the nature of the influence of the other variables needs to be explored further. Do the other variables magnify the effects of birth order, or do they cause the effect? Have positive results of birth order influence in other studies been caused by birth order or some of these other variables?

FOOTNOTE

1. Used for simplicity. Where significant on ratio data, analysis of variance are used subsequently.

REFERENCES

1. Bradley, R. W., "Birth Order and School Related Behavior: A Heuristic Review," Psychological Bulletin, 70:45-51, 1968.

2. Bradley, R. W.; Sanborn, M. P., "Ordinal Position of High School Students Identified by Their Teachers as Superior," Journal of Educational Psychology, 60:41-45, 1969.

3. Campbell, A. A., "A Study of the Personality Adjustments of Only and Intermediate Children," Journal of Genetic Psychology, 43:197-206, 1933.

4. McClure, R. F., "Birth Order and School Related Attitudes," Psychological Reports, 25:657-658, 1969.

5. Sampson, E. E., "The Study of Ordinal Position: Antecedents and Outcomes," in Maher, B. A. (ed.), Progress in Experimental Personality Research, Academic Press, New York, 1965, pp. 175-222.

6. Schacter, S., "Birth Order, Eminence, and Higher Education," American Sociological Review, 28:757-767, 1963.

PIAGET AND THE MENDE

OF SIERRA LEONE[1]

R. OGBONNA OHUCHE
University of Sierra Leone

ABSTRACT

Piaget's (4) principle of conservation of quantity has been demonstrated by many researchers to apply to American, European, and West African children. This investigation set out to determine the relevance of Piaget's ideas to a particular ethnic group in Sierra Leone. Mende pupils (N=231) of three age groups in sixteen schools located in towns or villages in four districts were interviewed on three tasks. The results showed that pupils in the 7 to 8 year age group received significantly higher ($p < .01$) scores than those younger than 7. It was concluded that conservation of quantity varies with chronological age and task and that there are individual differences.

PIAGET (4) claims that the notion of number is only gradually developed in children and that there are stages in this development. For him, the stages are stages in this development. For him, the stages allow for individual differences among children but there is a consistent general trend.

For the average European child used in Piaget's investigations the years before 7 are years in which the child's understanding of quantity concepts is dominated by his perception. He is, as it were, carried away by physical appearance. For him a number may be greater than itself, equal to itself, or less than itself depending on circumstances. However, from the age of 7, the child begins to appreciate that certain properties of objects remain unchanged in spite of external transformations. He attains what Piaget calls the conservation of quantity. In the words of Piaget:

Thought is no longer tied to particular states of the object but is obliged to follow successive changes with all their possible detours and reversals, and it no longer issues from a particular viewpoint of the subject, but coordinates all the different viewpoints in a system of objective reciprocities (5:8).

Additionally, Piaget's Geneva School of Thought has studied and continues to study the ability of children to relate sets of objects in one-to-one correspondence. Those children who understand the true equivalence of sets are able to relate. Involved in

this comprehension is a principle which Piaget calls cardination. Again, on the average, only children 7 years of age or older can apply this principle. Moreover, such children, unlike younger ones, are able to seriate or arrange objects into transitive asymmetrical sequences.

Many studies in Canada, the United Kingdom, the United States, and West Africa have used Piaget's theories with the development of quantity concepts as a springboard. Some of these have been reviewed by Almy (1), Flavell (2), and Ohuche (3). In the words of Almy:

On the whole . . . the bulk of the replication studies in the literature supports the notion that the child's ability to conserve quantity and number is arrived at gradually, and that the period of nonconservation, or perceptional domination is followed by a transitional stage, before conservation becomes pervasive. This conclusion, so far as we were able to determine from our review of literature, comes almost without exception from studies involving children of different ages (1:34).

The present study aimed to evaluate Piaget's position on conservation of quantity among children of the largest ethnic group in Sierra Leone, the Mende. Following the results of a pilot project which involved sixty Mende children in two strategically chosen elementary schools, a cross-sectional study including

most of Mendeland was designed. It looked at 231 Mende pupils in each of the four districts in which the Mende are the majority tribe: Bo, Kenema, Moyamba, and Pujehun.

METHOD AND PROCEDURE

Subjects

The design called for a sample of 240 pupils, fifteen in each of sixteen schools stratified in terms of district, location in chiefdom town or village, and the chronological age of the children. The schools in each district located in villages were arranged alphabetically and from each group two schools were chosen resulting in a total of eight. Similarly, eight schools were selected from those located in chiefdom towns. In each school, five children were selected randomly from each of the three age-group populations: 5 but less than 6, 6 but less than 7, and 7 but less than 8. All the requirements were satisfied by 231 pupils.

Tasks

The Ss were given an oral interview which involved testing for performance on three tasks. The sequence of tests on each task was preceded by a training session.

The training for Task 1 was designed to insure that each S could count up to ten in both English and Mende since ability to count to that level was essential for the test. Ten Star bottle tops were arranged in two equal rows for this purpose.

When it was clear that the child could count up to ten, the bottle tops were arranged in three rows of 3-4-3. The child was then asked to tell the number of bottle tops without counting. Next, the ten tops were arranged in one row and again the number of tops was asked for. Finally, the S was expected to give a reason for his or her answer.

This task involved the conservation of the number of two rows of bottle tops (that had been counted) through two transformations. A candidate could score 0 or 1 point on each of three questions. A total score of 2 or 3 was accepted to represent conservation.

Similarly, Task 2 incorporated the concept of the conservation of equality of two rows of bottle tops through two transformations. The idea was to compare the number of seven Star bottle tops with the number of seven Sprite bottle tops by using two different arrangements. Each candidate could score 0 or 1 on each of seven questions. A total score of 4 or more was considered to represent conservation.

The training for Task 3 was geared to making sure that the S understood the concept "heavier than" and therefore the meaning of the equality of two weights. Two equal match boxes were used. One was loaded with matches while the other was loaded with stones. Each S was made to feel the two weights and to answer probing questions. When the interviewer was satisfied that equality of weight was induced, the test was given. The issue here was to discover if the process of tying or untying a knot in a 3-inch shoe lace would cause the S to think that the

weight of an assemblage which included the shoe lace, a shallow container about 5 inches in diameter, and four match sticks changed. Allowable scores were 0 or 1 on each of three questions. A score of 2 or 3 was necessary for conservation.

RESULTS AND DISCUSSION

First, the data were analyzed qualitatively. A noticeable trend was the failure of the average less-than-7 year old to give good reasons for "yes" or "no" responses. The typical member of this set did not respond well to questions like "Why do you think so?" Other representative reactions from such children were: "The tied ones are heavier because they are tied." "The loose ones are heavier because they are spread out."

On the other hand, the typical child older than 7 invariably came up with a reason that made sense. Where the less-than-7 year old would start recounting ten bottle tops which were spread out in his presence after he had counted them, the typical pupil above 7 would respond "I know that there are ten because I had counted them and you have not removed any."

Second, the data were analyzed statistically. Table 1 shows the percentages of pupils of different age groups who verbalized ability to conserve. Bartlett's test of homogeneity indicated that the sample represented a homogeneous population.

A multiple classification analysis of variance was carried out (see Table 2). Three null hypotheses were used. The first was that there was no difference in mean scores due to district. The second was that village pupils obtained essentially the same scores, on the average, as pupils from chiefdom towns. The third was that there were no differences among the average scores of the three distinct age groups used in the study.

There was no basis for rejecting two of the three null hypotheses since the F values of 1.58 and 2.53 (see Table 2) were less than necessary for significance beyond the .05 level. Therefore, the corresponding variables (location in town or village and district) were not responsible for any differences in performance on the tasks. Moreover, none of the various interactions among variables were significant (see Table 2).

On the other hand, the variable age resulted in an F value of 154.05 (see Table 2). This F value was significant beyond the .01 level. Obviously, such a difference in performance could not be attributed to chance.

TABLE 1

PERCENTAGES OF PUPILS OF DIFFERENT AGE GROUPS WHO CONSERVED

TASK	AGE		
	5 but less than 6	6 but less than 7	7 but less than 8
One	36.0	52.6	92.0
Two	24.0	27.5	79.0
Three	6.5	14.0	52.0

TABLE 2

ANALYSIS OF VARIANCE FOR THE SCORES OF 231 PUPILS

Source	df	SS	MS	F
District	3	30.84	10.28	2.53
Town or Village	1	6.42	6.42	1.58
Age	2	1250.90	625.45	154.05*
Age x District	6	25.50	4.25	1.05
Age x Town or Village	3	78.26	39.13	9.64
District x Town or Village	3	20.32	6.77	1.67
Age x District x Town or Village	6	40.89	6.82	1.68
Within	207	841.27	4.06	
Total	230	2294.40		

*$p < .01$

This significant difference permitted the use of the Tukey procedure (6) for comparing means (see Table 3). Two of the differences, 5.06 and 3.70 (see Table 3), were greater than the computed difference of 2.22. Thus, the conclusions were drawn that the 7 but less-than-8 age group pupils were superior in their performance to both the 5 but less-than-6 age group and the 6 but less-than-7 age group.

TABLE 3

TUKEY COMPARISON OF MEANS

Age group	Mean	Mean-5.06	Mean-6.42
7 but less than 8	10.12	5.06	3.70
6 but less than 7	6.42	1.36	0.00
5 but less than 6	5.06	0.00	

Furthermore, this difference would appear not to have been due to difference in ability to verbalize since most Ss verbalized well during the tests. They communicated freely with the interviewers in Mende or English. It seems that these findings confirm the results of other investigators (1) that conservation abilities vary with age of the Ss and with tasks, and that there are individual differences.

FOOTNOTE

REFERENCES

1. Almy, Millie; Chittenden, Edward; Miller, Paula, Young Childrens Thinking, Teachers College Press, New York, 1966.

2. Flavell, J. H., The Developmental Psychology of Jean Piaget, D. Van Nostrand, Princeton, New Jersey, 1963.

3. Ohuche, R. O., "The Implications of Piaget's Results for the Mathematical Education of West African Children," Journal of Education, Ministry of Education, Freetown, 4:(no. 2)34-36, 1969.

4. Piaget, Jean, The Child's Conception of Number, Routledge and Kegan Paul, London, England, 1952.

5. Piaget, Jean, "Matter, Structure, and the Notion of Number," in Ripple, Richard E.; Rockcastle, Verue N. (eds.), Piaget Rediscovered—A Report of the Conference on Cognitive Studies in Education and Curriculum Development, School of Education, Cornell University, Ithaca, New York, 1964.

6. Snedecor, George W., Statistical Methods, The Iowa State University Press, Ames, 1956.

TEST ANXIETY AND DEFENSIVENESS EXPERIMENTALLY INDUCED BY FOUR CONDITIONS OF TESTING AROUSAL

BARTON B. PROGER, LESTER MANN, RAYMOND G. TAYLOR, JR.
Research and Information Services for Education, King of Prussia, Pennsylvania
and
JAMES E. MORRELL
Centennial School District, Southampton, Pennsylvania

ABSTRACT

To study the relationships among frequency of testing, arithmetic learning and retention, predispositional test anxiety, defensiveness against admission of test anxiety, and induced test anxiety, eighty sixth-grade Ss were randomly assigned to four arousal conditions: tests every day, tests every other day, tests once a week, and daily practice. Teachers were randomly rotated daily. The study lasted 5 weeks. An achievement posttest was given at the end of the study and again 2 weeks later. Induced test anxiety was measured at the end of each week. On both achievement posttests, the only significant difference was in favor of the daily test group over the weekly test group. Induced test anxiety was found to operate similar to predispositional test anxiety.

TEST ANXIETY in realistic classroom settings has usually been studied in a predispositional sense (8). That is, Ss are given a measure of test anxiety before any treatments are applied in order to classify them into predispositionally anxious groups, such as high, medium, and low. One or more tasks are then given to Ss to see how the predispositional levels of test anxiety relate to ongoing performance or achievement. However, Ruebush (7) has also discussed how experimentally manipulatable or induced test anxiety can be studied in relation to performance. Here, Ss can be randomly assigned to various treatment groups, stress-producing treatments of various degrees can be applied, and then a measure of test anxiety can be given to detect whether or not such affect has been induced in them. Because of the administrative complexities involved, induced test anxiety studies have rarely been conducted in realistic classroom testing situations (4).

In a pilot study (4) ninety-three third-grade arithmetic pupils were randomly assigned to two treatments of high predictability: daily testing or daily practice. Methods, sex, and IQ (high, medium, and low) were subjected to analysis of variance. An immediate arithmetic achievement post-test was given at the end of the 4-week study and again 3 months later. To measure any affects that might be induced by the treatments, the Test Anxiety Scale for Children (TASC) and the Defensiveness Scale for Children (DSC) were given at the close of the study. The treatments produced comparably high levels of anxiety, defensiveness, and achievement in all Ss. One possible explanation of the results is that since the daily testing and the daily practice stress conditions were both highly predictable by Ss after a few days, comparable behavior in the affective and cognitive realms was to be expected.

To investigate further the nature of experimentally induced test anxiety, defensiveness, and achievement under realistic learning conditions, the present study was conducted with stress-producing treatments of both high and low predictability: daily testing and daily practice (highly predictable by Ss) and testing every other day and testing only once a week (less predictable by Ss). On the basis of past research (4), it was hypothesized that the two highly predictable stressful conditions would produce significantly lower anxiety and defensiveness scores than the two less predictable treatments. These hypotheses were predicated upon the arousal theory of Hebb (2).

PROCEDURE

The 5-week experiment was conducted in the late spring of 1968. The Ss (working N = 80) consisted of the entire sixth grade of a suburban elementary school in a middle-class community in the Greater Philadelphia Area. The pupils represented an average to above-average ability in composition (average fifth-grade, Lorge-Thorndike IQ = 115.00).

There were three intact arithmetic classes. Each Monday, Wednesday, and Friday, these three classes received their usual arithmetic instruction for an hour following recess in the latter part of the morning. On Tuesdays and Thursdays these classes met for 45 minutes. To control differential practice effects from taking or not taking the experimental tests, the testing process was isolated from the usual instructional cycle in arithmetic by having Ss receive tests or equivalent practice early in the morning each day. The test or practice was on the previous day's work. To control any effects due to differences in presentation of new material among the three regular intact classes, Ss of the three stress groups were randomly assigned to the four stress-producing conditions of the experimental testing periods. To control differential effects of teacher personalities and efficacies, a random rotation schedule of teachers was used for the testing period. The original three arithmetic teachers were used, and the most competent student teacher became the fourth proctor. No tests or practice were given on the last day of each school week, which was reserved for an abbreviated version of TASC (5).

To control further the differential practice effects that might otherwise be caused by varying test contents on any given day, the senior investigator devised an identical worksheet format for both test and practice groups during the experimental testing period. Depending upon the schedules of testing within the treatment groups, each of the four groups was told that the worksheet was a test or was only practice, as the case might be, and the headings on the worksheets reflected these facts accordingly.

A specially devised immediate achievement posttest (46 items, with two items of each type of fraction and division problem covered during the 5-week unit of instruction; see Proger (5)) was given, and the same test was given as a measure of delayed achievement 2 weeks later. The immediate achievement posttest had a Spearman-Brown-corrected, odd-even coefficient of internal consistency of 0.97. To obtain measures on the predispositional levels of test anxiety on the first day of the study, the unabridged versions of both the commonly accepted TASC and the Defensiveness [against admission of test anxiety] Scale for Children (DSC) were given. In this study, the 30-item TASC had a split-half reliability of 0.85 (corrected), and the reliability of the DSC calculated the same way was 0.76. At the end of each school week, an abbreviated test anxiety scale (the 12 items from TASC judged to be most pertinent to induced test anxiety; see Proger (5)) was given. For the first, third, and fifth administrations of the 12-item scale, the respective split-half coefficients of internal consistency were 0.78, 0.80, and 0.83.

RESULTS

The three analyses of variance (the two achievement posttest analyses and the induced test anxiety trials analysis) will be considered first. All factors in each analysis were considered fixed because of the specific factor compositions. The cell frequencies of the data matrices were unequal because of random experimental mortality. Therefore, the unweighted-means approach was deemed more appropriate than the least-squares, unequal-frequency method (10:374). The results of the two achievement posttest analyses of variance are presented in Tables 1 and 2.

On the three-way analyses of variance of both posttests, the four testing methods were significantly different at the .05 level. On both posttests, the control factor of previous achievement operated quite effectively at the .05 level in removing variance from the error component. In neither analysis were the main effect of sex or the interaction effects significant. Scheffe's technique for multiple comparisons (1) showed that the daily test procedure was significantly more effective than the weekly test approach at the more conservative .05 level on both posttests; no other individual comparisons were significant. The means of the four groups on the 46-item immediate achievement posttest were: daily tests, 33.05; weekly tests, 24.63; alternate days, 22.57; and daily practice, 26.82. The means of the groups on the same test given as a measure of retention were: daily tests, 31.35; alternate days, 23.11, weekly tests, 20.41; and daily practice, 24.56.

The third unweighted-means analysis of variance to be considered is that of the induced test anxiety trial means. Because previous work in the area of test anxiety induced by different schedules

TABLE 1

UNWEIGHTED-MEANS ANALYSIS OF VARIANCE: IMMEDIATE ACHIEVEMENT POSTTEST

Source of Variation	Sum of Squares	d f	Mean Square	F Test
Methods (A)	892.70	3	297.57	3.58 *
Sex (B)	191.19	1	191.19	2.30
Previous Achievement (C)	3191.29	1	3191.29	38.45 **
A x B	51.70	3	17.23	<1
A x C	35.95	3	11.98	<1
B x C	0.12	1	0.12	<1
A x B x C	26.09	3	8.70	<1
Within Cell	5140.96	62	82.92	

*.01 < p < .025
**p < .005

TABLE 2

UNWEIGHTED-MEANS ANALYSIS OF VARIANCE:
DELAYED ACHIEVEMENT POSTTEST

Source of Variation	Sum of Squares	d f	Mean Square	F Test
Methods (A)	1017.57	3	339.19	3.74*
Sex (B)	270.26	1	270.26	2.98
Previous Achievement (C)	3415.01	1	3415.01	37.65**
A x B	8.68	3	2.89	<1
A x C	15.97	3	5.32	<1
B x C	0.36	1	0.36	<1
A x B x C	24.91	3	8.30	<1
Within Cell	5714.98	63	90.71	

* $.01 < p < .025$
** $p < .005$

of testing yielded no significant differences on either sex or previous achievement (4), the factorial model chosen to study induced test anxiety dealt only with methods and trials (see Proger (5) for comparable results yielded by the model with the two additional factors of sex and previous achievement in it). To accommodate the unequal cell frequencies, the unweighted-means, repeated-measures design of Winer (10:376-378) was used. The results of this analysis are presented in Table 3.

TABLE 3

UNWEIGHTED-MEANS ANALYSIS OF VARIANCE:
INDUCED TEST ANXIETY

Source of Variation	Sum of Squares	d f	Mean Square	F Test
Between:		83:		
Methods (A)	50.02	3	16.67	<1
Ss Within Groups	2701.28	80	33.77	
Within:		336:		
Trials (B)	99.63	4	24.91	15.87*
A x B	9.21	12	0.77	<1
B x Ss Within Groups	501.72	320	1.57	

* $p < .005$

Only the main effect of trials (abbreviated test anxiety scale administrations) was significant at the .05 level. The weekly overall induced test anxiety averages were: trial 1, 4.02; trial 2, 3.38; trial 3, 3.27; trial 4, 2.70; and trial 5, 2.73. The greatest decrease in anxiety occurred between the first and second administrations (p < .005). Finally, since there was no significant interaction, only the overall trends were calculated as in Winer (10:273-275). There was a highly significant linear decrease at the .05 level (F = 57.20, p<.005), but the quadratic, cubic, and quartic trends were all insignificant.

The last analysis undertaken was to study the interrelationships among predispositional defensiveness against admission of test anxiety, induced test anxiety, and immediate posttest achievement. The product-moment coefficients of correlation are given in Table 4.

The matter of induced test anxiety will be considered first. Contrary to hypothesis, the two stress-producing treatments most predictable by Ss (daily testing and daily practice) did not produce significantly lower anxiety and defensiveness levels. All four treatments induced comparable degrees of anxiety and defensiveness. Further, all four treatments demonstrated comparable decreases in manipulatable anxiety throughout the study. One explanation for the decrease in test anxiety from trial to trial is that the continued testing during the experiment allowed Ss to become accustomed to frequent tests. This viewpoint holds some validity to the extent that the daily practice-feedback condition can be considered as an evaluative situation. Perhaps a more plausible explanation is that Ss became desensitized to the short, 12-item anxiety scale itself and thus answered more and more casually as the experiment progressed. Some evidence supporting this second explanation was provided by the teachers themselves—they noted that the pupils vocally expressed their displeasure and boredom with answering "the same old questions" on test anxiety each week of the study.

A final note on the weekly decreases of induced test anxiety seems in order. The largest decrease occurred between the first and second administrations of the abbreviated TASC. In weekly retesting with a standardized achievement test, Mann and others (3) found that Ss improved from trial to trial, and that the largest change occurred between the first and second trials. The measurement of periodic change has always been a delicate matter, fraught with difficulties. For cognitive retesting, one might have to deal with test-wiseness. In this study, for affective retesting, instrument desensitization seems to be most pertinent. However, the methodology of repeated testing in the field of emotionality is not yet well enough understood to provide detailed explanations.

Further inferences about the nature of induced test anxiety as it operated in this study can be drawn from the intercorrelations in Table 4. As would be expected on the basis of past studies at the elementary school level (8), the correlation of -0.19 between immediate posttest achievement and predispositional test anxiety is significant at the .05

TABLE 4

RELATIONSHIPS AMONG IMMEDIATE POSTTEST ACHIEVEMENT (I.A.), PREDISPOSITIONAL TEST ANXIETY (TASC), PREDISPOSITIONAL DEFENSIVENESS (DSC), AND INDUCED TEST ANXIETY (A-TASC 1, A-TASC 3, AND A-TASC 5)

	I.A.	TASC	DSC	A-TASC 1	A-TASC 3	A-TASC 5
I.A.	1.00	-0.19	-0.002	-0.16	-0.15	-0.30
TASC	—	1.00	0.31	0.72	0.60	0.54
DSC	—	—	1.00	0.21	0.16	0.13
A-TASC 1	—	—	—	1.00	0.78	0.76
A-TASC 3	—	—	—	—	1.00	0.90
A-TASC 5	—	—	—	—	—	1.00

*N=80; significance was determined by t-test (9:281), testing H_0: r = 0.00 against the appropriate one-sided alternative (on the basis of past predispositional test anxiety and predispositional defensiveness studies). For N=80, the critical values of r and the corresponding levels of significance are: 0.36 (p=.0005), 0.29 (p= 0.005), 0.26 (p=0.01), 0.22 (p=0.025), 0.19 (p=0.05), 0.15 (p=0.10), and 0.10 (p=.20).

level. Similarly, as found in the third-grade frequent testing study of Mann and others (4), the r = -0.002 between immediate posttest achievement and predispositional defensiveness is insignificant (p> .20). A third result that supports the usual expectations under a predispositional test anxiety framework is the r = 0.31 between TASC and DSC (p< 0.005). However, to the investigators' knowledge, no studies to date have related levels of predispositional anxiety to levels of induced anxiety under realistic, longitudinal learning conditions. The correlations of every other week's induced test anxiety levels, as measured by the abbreviated anxiety levels (A-TASC 1, A-TASC 3, A-TASC 5), with TASC (A-TASC 1, A-TASC 3, A-TASC 5), with the original predispositional anxiety levels, as measured by TASC, are also given in Table 4. In general, there is a relatively high relationship between induced and predispositional anxiety.

Finally, the induced achievement results will be discussed. In summary, it was found that only the program of daily testing resulted in significantly higher achievement at the .05 level in both posttests. The alternate day test group, the weekly test group, and the daily practice group were not significantly different at the .05 level on either posttest. A graphical summary of the induced achievement results is shown in Figure 1. One should note that the arrangement of methods groups on the horizontal axis in order of increasing amount of testing does not imply equal intervals or degrees of testing.

In our present state of knowledge, explanations of the psychological mechanism which operated to produce the achievement curves in Figure 1 must remain only conjecture. The following discussion is in terms of physiological and affective arousal phenomena, as might be inferred from Hebb (2). First, one must remember that each of the four groups in the present study was told on the day the worksheets were distributed whether the worksheets

FIGURE 1

COMPARISONS OF ACHIEVEMENT ON IMMEDIATE AND DELAYED POSTTESTS FOR FOUR METHODS GROUPS IN ORDER OF INCREASING DEGREES OF TESTING, POOLED ACROSS SEX AND PREVIOUS ACHIEVEMENT
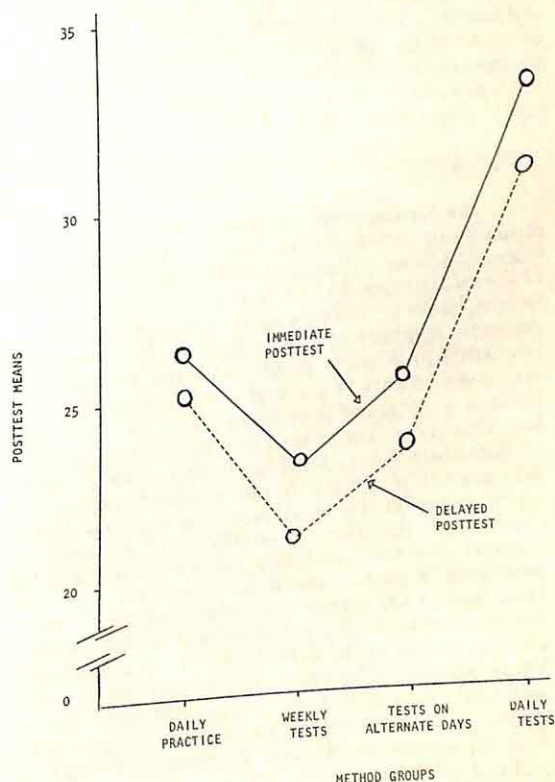
were to be counted as practice problems or as a test. The exact schedule of testing that any particular pupil was under was not made known to him at any point in the experiment. This procedure was necessary to control test expectation throughout all four groups. It is likely that after the first few days of the 5-week experiment, the pupils in the daily test group and the daily practice group inferred that they would be under a continuous schedule of tests or practice, respectively. Thus, while all four groups might have been under some form of emotional stress with respect to expectation the first few days, the daily test group and the daily practice group probably attained an optimal expectation arousal state. On the other hand, the alternate day test group and the weekly test group were under less obvious schedules of testing; the likelihood that these pupils could infer the schedule of testing they were under was perhaps decreased because of the interruptions by the 2-day weekends and the administrations of the test anxiety scale on the last school day of each week. Perhaps the least obvious schedule of testing was that of the weekly test group. This may account for the poorest performance of this group out of all four groups.

It may be, using the arousal (or drive activation) theory of Hebb (2:249-250), that the pupils in both the alternate day test group and the weekly test group were aroused to an excessive state of activation (with an unstable expectation state perhaps playing a crucial role in bringing about this arousal); thus, they could not perform as well as they might under more predictable conditions; these Ss might be subject to "increasing emotional disturbance, [and] anxiety" (2:250). On the other hand, the Ss in the groups undergoing the relatively stable expectation condition (the daily test group and the daily practice group) probably had not exceeded their optimal arousal level (that which produced the best performance) and hence had not yet lost as much relative "alertness, interest [and] positive emotion" (2:250).

IMPLICATIONS

The formal study of induced affect, as manipulated by different treatments in realistic settings, has been a neglected area. The methodology of the measurement of change in affect over time is complicated. With repeated administrations of the same measure of affect, there is the usual contamination problem of prior measures on successive ones. There is the much less commonly recognized problem of purging of affect, which all too often gives rise to the spurious conclusion of no difference among treatments with respect to induced levels of affect. In this second problem, after termination of a relatively long application of treatments, the view can be offered that any immediate posttests of cognition (given along with the measures of affect) signal to the subject the obvious end to any stress-reducing treatments; thus, induced levels of affect revert back to their usual chronic levels by the time they are measured. While one cannot be sure which of the two problems was present in this study, it is suggested that more emphasis be given to in-process measures of affect in longitudinal studies, as compared to pre- and post-measures.

Finally, one can ask what might be done in future research on the cognitive aspects of the stress-producing treatments used in this study. For example, the daily test group of the present experiment has shown that tests in and of themselves at the elementary school level can teach Ss material above and beyond only practice situations; the problem then arises as to exactly how such content learning arousal effects take place. Perhaps adopting the methods used by the "test-like event" investigators would enable one to attack this problem. Basically, "test-like events" are study-guide questions inserted into reading passages or assignments when given in class (that is, this approximates a test situation in its evaluative aspects as compared to study guide questions given as outside-class homework where the study situation is relatively informal and non-test-like). For example, Rothkopf and Bisbicos (6) have dealt essentially with completion-type review questions inserted into the text itself. The advantage in such procedures is that one can gain a great deal of control in studying an effect such as content structuring in "test-like" situations, that was hitherto unavailable.

FOOTNOTES

REFERENCES

1.   Edwards, Allan L., *Experimental Design in Psychological Research* (Rev. Ed.), Holt, Rinehart, and Winston, New York, 1960.

2. Hebb, D. O., "Drives and the C.N.S. (Conceptual Nervous System)," Psychological Review, 62:243-254, 1955.

3. Mann, Lester; Taylor, Raymond G. Jr.; Proger, Barton B.; Dungan, Roy H.; Tidey, William J., "The Effect of Serial Retesting on the Relative Performance of High- and Low-Test Anxious Seventh-Grade Students," Journal of Educational Measurement, 7:97-104, 1970.

4. Mann, Lester; Taylor, Raymond G. Jr.; Proger, Barton B.; Morrell, James E., "Test Anxiety and Defensiveness Against Admission of Test Anxiety Induced by Frequent Testing," Psychological Reports, 23:1283-1286, 1968.

5. Proger, Barton B., "The Relationship Between Four Testing Programs and the Resultant Achievement and Test Anxiety Levels of High- and Low- Previous Achievement Sixth-Grade Arithmetic Students," unpublished doctoral dissertation, Lehigh University, Bethlehem, Pennsylvania, 1968.

6. Rothkopf, Ernst Z.; Bisbicos, Ethel E. "Selective Facilitative Effects of Interspersed Questions on Learning from Written Materials," Journal of Educational Psychology, 58:56-61, 1967.

7. Ruebush, Britton K., "Anxiety," In Stevenson, H. W.; Kogan, J.; Spiker, C. (eds.), Child Psychology, Sixty-Second Yearbook, Part 1, National Society for the Study of Education, Chicago, 1963, pp. 460-516.

8. Sarason, Seymour B., Davidson, Kenneth S.; Lighthall, Frederick F.; Waite, Richard R.; Ruebush, Britton K., Anxiety in Elementary School Children: A Report of Research, Wiley, New York, 1960.

9. Tate, Merle W., Statistics in Education and Psychology: A First Course, Macmillan, New York, 1965.

10. Winer, B. J., Statistical Principles in Experimental Design, McGraw-Hill, New York, 1962.

THE DESIGN OF CORRELATION STUDIES

KEITH F. PUNCH
The Ontario Institute for Studies in Education

ABSTRACT

Correlation techniques will continue to be widely used, because of the ex post facto nature of much educational research. Within this constraint, the general aim of moving from description to the building of explanatory theory, and the operational aim of accounting for variance, point to the need for multivariate rather than bivariate correlation studies. This in turn points to multiple linear regression as the basic design and analysis tool, the more so in view of its power to answer as well typical analysis of variance questions. If, in isolated cases, bivariate correlation studies are unavoidable, they should systematically prepare for later multivariate studies. The error variance of measures used should in all cases be estimated and reported, given the aim of accounting for variance.

MUCH EMPIRICAL research in education, and in social science generally, is necessarily ex post facto in nature. This, in the general case, means correlation studies, and explains the widespread use of correlation techniques in reported research. It is not necessary to go into the usual laments about the deficiencies of ex post facto research and correlation studies. There are, of course, problems. But at least such studies, by not manipulating variables, admit a truer, less artificial view of the world of interest than do studies in the classical experimental design. What is needed, for better ex post facto research, is a set of techniques for handling complex multiple relationships, and logical checks and balances in the interpretation of relationships observed. Given the importance of correlation techniques, then, this article deals with some of the problems and strategies involved in developing and using techniques for handling complex multiple relationships. Though the points raised have design implications, the discussion is, for simplicity, cast mostly in data analysis terms.

The usual problem in data analysis is to obtain maximum information from the data, while not bruising them unfairly in quest of support for hypotheses. Here, by contrast, the reverse kind of problem is used as a starting point—a too simplistic interpretation of data, with the attendant dangers of invalid inference. The main purpose in many reported research studies is to establish a relationship between two variables. One of these, implicity or explicitly in the investigator's conceptualizing, is seen as "dependent" or "criterion," the other as "independent" or "predictor." Usually, though with varying degrees of plausibility, the data are assumed to meet certain conditions, and the product-moment coefficient, r, is used. The rules regarding the use of this coefficient are well documented, if not always well followed. Thus, for example, predictor variable x and criterion variable y are found to correlate $+0.3$, which, with a sample size of, say, 100, is significant beyond the .01 level. This permits strong confidence in the rejection of the null hypothesis $\rho = 0$, and in the acceptance of its alternative $\rho > 0$. The report then typically moves to a discussion of the implications of the relationship just established.

Now there is nothing wrong with this procedure in itself. By not going far enough, however, it enhances the possibility of invalid interpretation and inference. The point is that while r is a useful summary index of the relationship, r^2 —the square of the correlation coefficient—gives an estimate of that part of the variance in y held in common with, or accounted for by, variance in x. This point has been well documented in the literature (4), including its mathematical proof (3). In the above hypothetical example, then, an r of $+0.3$ has almost certainly not come about through sampling error, and indicates a real relationship in the population

under consideration. Yet with $r^2 = .09$, a mere 9 percent of the variance in y is accounted for by variance in x. If small or zero error variance in y is assumed, this leaves some 90 percent of its variance untapped. This point is too often unrecognized, and hence is omitted.

The possibility of invalid inference is clear, over and above any "correlation-causation" problem. That is, invalid inference is likely even if we establish, or assume, a casual influence from x to y. While x and y are clearly related, it is idle to consider changing x as a strategy for changing y. Yet this may well be the interpretation given or taken, especially in practitioner oriented reporting. It is the more likely when, as is usual, the report underlines the statistical significance of the observed r and stresses and discusses the relationship, but neither mentions nor interprets the r^2. In the above example, we really know very little at this stage about what varies with y, casually or not, given that 90 percent of its variance is uninvestigated.

For theory oriented research, the problem is just as bad. The relationship does hold, but the overall pattern of variation in y remains virtually unknown. Yet this is surely one of the basic research problems. Ultimately, research seeks answers to questions of the kind "What causes y?" For reasons of strategy, this is rephrased as "What causes y to vary?" To avoid questions of causation, this in turn is rephrased as "What is the pattern of variation of y?" That is, what variables are associated with y, or hold variance in common with y? In other words, research seeks to account for as close to 100 percent of the variance in the y under consideration as is possible. Seen in this light, and acknowledging multiple causation, to delineate isolated 2-variable relationships is to make only the first research step.

In substance, then, while r may be highly significant statistically, r^2 may be very small. Indeed, with large enough sample size, a correlation of .01 reaches statistical significance, but the proportion of variance accounted for is negligible. In the face of this problem, two strategies are possible when it comes to designing research. Which should be used depends on the amount already known in the area of concern. The first applies when very little can be claimed in advance about variation in the y under study. The second applies when the influence of variables other than the x on any particular y is known or suspected. Both keep in mind that accounting for variance is more the goal than testing for isolated relationships.

Consider the case where we wish to investigate the relationship between any particular x and y. We suspect that, while the relationship will hold, x will at best account for only a small proportion of the variance in y. At the same time, we cannot reasonably propose what variables other than x will be significant correlates of y. That is, we are faced either with something like "random speculation" to identify these other variables, or with the mechanical use of standard, easy-to-collect, base line data (sex, age, IQ, socio-economic status, etc., variables used far too

often and far too mechanically). Neither alternative belongs in a study designed for economy of, and maximum output from, data collection. In this case, the study should be programmed so that the x-y relationship is investigated as a first step, but the researcher returns to the sample, on the basis of the correlation results, to search systematically for the other correlates of y.

The rationale is simple. Assume that the analysis produces the hypothetical figures used above ($r_{xy} = +0.3$, N = 100). It is now possible to rank-order the one hundred Ss according to their standard scores on variable x, and by dividing them at the median, to designate fifty as high and fifty as low. Independently of this designation, the hundred Ss can likewise be ranked according to standard scores on variable y, with fifty designated high and fifty low. Each S now has a 2-way classification, either (1) high-high, (2) high-low, (3) low-high, or (4) low-low. Clearly, it is groups 1 and 4 which bring about whatever positive correlation exists, and groups 2 and 3 which work against it. Two groups have now been identified—those who contribute to the correlation, and those who detract from it. The researcher can now return to the sample to search systematically for differences between the two groups. If only groups 1 and 4 were used, the computed correlation would be stronger, and positive; if only groups 2 and 3 were used, it would be weaker and perhaps negative. This suggests the question: what differentiates Ss in 1 and 4 as a group, from those in 2 and 3 as a group? To answer this question is to indentify other potentially important predictor variables, of whatever kind. The next study in this area can then incorporate these variables into the investigation for greater payoff, in terms of the proportion of variance in y "explained."

The obvious difficulty here is the practival one of returning to the sample. Yet this kind of analysis would increase the effectiveness of bivariate correlation studies. Since such studies represent only a small first research step, they ought to attempt to provide direction for later multivariate studies in the particular area. Clearly, the above procedure can be modified where necessary. Thus, for example, finer classification schemes could be used with each variable—very high, high, medium, low, very low, perhaps. This would both give more clearly defined groups in terms of those which contribute to the relationship and those which do not, and reduce the number of Ss to be reinvestigated. Whatever the classification scheme used, the point is that some such procedure would increase the dividends from a piece of research necessarily limited to investigating a 2-variable relationship.

But all of this may not be necessary. Omitting practical problems of data collection, how often must research, because of insufficient knowledge, consider the relationship between only two variables, between only one predictor x and the criterion y? At least not very often in educational research, and perhaps never, for two reasons. Firstly, in any area there is a great deal of knowledge and near-knowledge, albeit fragmented and uncertain. The researcher must systematically exploit all that is known relevant to his problem. Further, and sec-

ondly, he must adequately analyze his problem. An orientation toward explanation through theory building as the goal of scientific research, rather than only description and documentation, is essential here. Thus an hypothesized relationship between x and y should be built on some basis, and that basis should be exposed and analyzed. The question then becomes: Why should this relationship exist, if it in fact does? Research too often remains at the level of mere description, whereas the goal of science is the building of explanatory theory. Given a tight, logical structure of the type "If the theory is true, then the hypothesized x-y relationship follows," testing the hypothesis by documenting the relationship becomes the vehicle for confirming or rejecting the theory.[1] An adequate explanation for the hypothesis will normally involve (a) specifying the intervening variables by which x is connected to y, (b) identifying the conditions under which x will be more or less strongly related to y, and sometimes (c) showing the proposed relationship to be an instance of some wider generalization. Points a and b force consideration of the other possible predictors of y, either in conjunction with x or independent of it. An adequate analysis would indicate whether they are to be seen as additional predictors, or as control variables. With adequate operationalization, they can then be incorporated into the research.

The point then is that proper problem analysis, and attention to theory building prior to structuring and gathering data will convert most bivariate studies into multivariate ones. Multivariate studies require multivariate techniques. Assume we now have one criterion variable y, and a series of predictor variables $x^{(1)}$ $x^{(2)}$ .... $x^{(n)}$. For such cases, multiple linear regression is the general, basic design and analysis tool. An exposition of the technique is available from numerous sources (1, for example). For present purposes, we should note that three important kinds of questions can be answered.

(i) We can estimate the proportion of variance in y accounted for by, or held in common with, all predictors $x^{(1)}$ $x^{(2)}$ ... $x^{(n)}$, considered together. Just as one predictor variable study yields r and more importantly r^2, so multiple predictor studies yield R, the multiple correlation, coefficient, and, more, importantly, its square R^2. The R^2, of course, will be between 0 and 1, and its statistical significance is testable through F. It represents the most direct answer to the question of accounting for variance, proposed earlier as one of the basic research questions.

(ii) Within the limits of the answer to (i), we can begin to assess the relative order of importance among the predictors in accounting for variance. There are at least two ways of doing this. One is by adding (or deleting) predictor variables in prediction systems using stepwise regression procedures, to determine the contribution to prediction of the variable added or deleted over and above that of other predictors.[2] The problem here is that with interrelated predictors, the relative importance of each will likely depend on the order in which it is entered into, or dropped from the regression model. No statistical solution exists to the problem of "the correct" order for entering predictors into a regression model. The researcher may, however, be able to justify a conceptual (or perhaps, temporal) ordering among the predictors. If so, or in the unlikely event of unrelated predictors, the stepwise routine can give the proportion of variance associated with each predictor, within the cumulative limits of (i). The other approach is through the standardized partial regression coefficients, or beta weights. If for example the beta weight for predictor $x^{(1)}$ is .45, and that for $x^{(2)}$ is .06, and the difference between these weights is significant, $x^{(1)}$ is more important a determinant of y than is $x^{(2)}$. This is so because a beta weight indicates how much change in y is produced by a standardized change in the particular predictor variable with other predictors held constant. Which of the two ways to use should depend on the interpretation given to "the relative importance of predictors." A precise operational version of this phrase would, in most cases, point to the second and neater method, the beta weight analysis.

(iii) We can test hypotheses of the kind normally tested only by analysis of variance and analysis of covariance procedures. In general, such hypotheses will deal with differences between groups on the criterion, where the sample units are assigned to groups according to some univariate or multivariate classification system. It has only recently been shown that analysis of variance and of covariance may be seen as special cases of multiple linear regression analysis (2). It follows then that the more general technique will be more useful, especially since it appears that multiple regression is easier for the average researcher to use and apply than are the traditional procedures. Thus hypotheses dealing with between—group differences can be tested within the context of the important questions of (i) and (ii) using regression analysis. Furthermore, both continuous and discontinuous predictor variables can be handled in the same analysis. And, in the process, the scalability of "doubtful" variables can be easily determined using the hypothesis testing function of regression analysis. This may be a most important payoff, given the uncritical acceptance of the assumptions involved in ordinal measurement, and hence the large number of educational research variables whose scalability is doubtful.

All three types of question are important, and relatively easy to answer using multiple linear regression analysis. And because it can be used to assess between group differences on the criterion—whether univariate or multivariate classification determines the groups, and whether or not covariates are considered—regression analysis appeals

as the appropriate analysis tool both for ex post facto research and for research using the less typical experimental or quasi-experimental designs. The rationale behind the technique is easily understood, and the technique permits ready communication with the computer. Its use will reduce reliance, in both design and analysis, on traditional, often over-rigid statistical formulas and procedures. Further, as has been suggested, an orientation toward regression analysis as the mainstream tool will assist in transforming most bivariate studies into multivariate ones. Not that this orientation itself will indicate which other variables to consider. That should be dictated by problem analysis, and by previous research and knowledge, integrated by the researcher into a unified, consistent, and clear theoretical framework.

One further point should be mentioned in this general discussion of the design of correlation studies. It is important to know the probable amount of error variance in all variables the study measures, but particularly in the criterion variable. Error variance is, by definition, spurious, random or untrue variance—that is, variance which cannot be accounted for. It is a basic research problem, to repeat, to account for as much of the variance as possible in the particular criterion variable of interest, using its relationships with predictors. If criterion error variance is high, there is clearly something less than 100 percent of its variance to be accounted for. Error variance is estimated by the use of reliability coefficients, however computed. This, then, is one reason why reliabilities for all measurements used in research should be estimated and reported, especially for those measurements which represent the operational version of the criterion variable of interest.

FOOTNOTES

1. The term "theory" only means here a set of logically consistent and plausible propositions which serve to explain, in an if-then sense, the hypothesis. With the if-then set up, it is clear that theories are never proved this way, because of the logical fallacy of "affirming the consequent."

2. The rationale is simple: F is used to test the difference between the R^2 from the model with the predictor, and that from the model without the predictor, with other predictors in both models.

REFERENCES

1. Baggaley, A. R., Intermediate Correlational Methods, John Wiley, New York, 1964.

2. Bottenberg, R. A.; Ward, J. H., Jr., Applied Multiple Linear Regression, Technical Documentary Report, PRL-TDR-63-6, Lackland Air Force Base, Texas, 1963.

3. Kelly, F. J.; Beggs, D. L.; McNeil, K. A.; Eichelberger T.; Lyon, J., Research Design in the Behavioral Sciences: Multiple Regression Approach. Southern Illinois University Press, Carbondale, 1969.

4. McNemar, Q., Psychological Statistics, John Wiley, New York, 1962.

SOCIAL CLASS, OCCUPATIONAL ASPIRATION,

AND OTHER VARIABLES[1]

M. S. TSENG[2]
West Virginia University

ABSTRACT

Measures of occupational aspiration, perception of occupational prestige, achievement motivation, and fear of failure of 179 high school boys were investigated with Ss' socioeconomic and grade levels as independent variables. ANOVA on a 3x4 factorial design and multiple comparisons showed that Ss from the lower and lower-lower socioeconomic groups had significantly lower occupational aspiration and more distorted perception of occupational prestige hierarchy than Ss from the middle class and that among the ninth grade Ss, the middle-class group possessed significantly higher achievement motivation than the lower and lower-lower groups, whereas among Ss from the lower-lower class, the twelfth grade group showed significantly higher achievement motivation than the ninth and tenth grade groups. Analyses of covariance were also carried out on a 3x4 factorial design with regard to occupational aspiration and perception of occupational prestige and the four dependent variables were factor analyzed and results discussed.

THAT CERTAIN personality variables relate to socioeconomic status has long been recognized by social scientists. In earlier work by the National Opinion Research Center (7), it was established that perception of the occupational prestige hierarchy was positively related to socioeconomic class. Beilin (3) reported that there was an availability of high level talent in the lower socioeconomic groups but problems existed in developing this talent because many lower-class individuals did not attempt to get the necessary education or training needed for high level jobs. Tseng and Thompson found that students of lower class tended to select lower level occupations (9), and that significantly fewer students from the lower class sought counseling (8).

The primary purpose of the current study was to investigate whether male high school students of various socioeconomic groups differ significantly in occupational aspiration, perception of the occupational prestige hierarchy, achievement motivation, and fear of failure.

Occupational aspiration is the degree to which the individual aspires to set as a goal the attainment of occupations of certain prestige level, whereas perception of occupational prestige indicates the strength of affective (positive or negative) response he attaches to occupations. Achievement motivation (or need achievement) is the degree of competitiveness for excellence present in a given individual which is viewed as the motive to approach success, whereas fear of failure or anxiety level aroused by the success-failure cues is considered as the motive to avoid failure.

Specific questions examined in the study were as follows:

1. What are the similarities and differences among the middle, lower, and lower-lower socioeconomic groups with regard to occupational aspiration, perception of occupational prestige, achievement motivation, and fear of failure?

2. What are the similarities and differences among the ninth, tenth, eleventh, and twelfth grade male students in terms of the four variables?

3. Is there a significant interaction between the socioeconomic status and grade level in relation to the four dependent measures?

SUBJECTS

The population, consisting of 5,600 persons, was defined as all ninth, tenth, eleventh, and twelfth

grade male students and male drop-outs who dropped out during the school year 1966-67 and were permanent residents of McDowell County, West Virginia.

Out of a sample drawn at random, 179 boys who were non-drop-outs and provided all the data including the grade level and social class data as well as the four dependent measures were selected as the Ss of this study.

Of this group of 179, there were twenty-nine, fifty-four, forty-seven, and forty-nine students enrolled in the ninth, tenth, eleventh, and twelfth grade, respectively. Among these subjects fifty-eight were in the middle class, sixty-four were in the lower class, and fifty-seven were in the lower-lower class.

The classification of socioeconomic level was made on the basis of father's occupation and father's and mother's educational level. National Opinion Research Center (NORC) scores of fathers' occupations ranging from 1 to 49 were classified as being in the middle class, from 50 to 76 belonged to the lower class, and from 77 to 90 belonged to the lower-lower class. The NORC score was obtained by assigning the ranking of the occupation in terms of its prestige level as classified by the National Opinion Research Center (7) with, for example, 2 representing physician, 10 representing banker, 60 representing plumber, and so forth. The smaller the value of the NORC score the higher the level of prestige of the occupation. The cutoff points for father's and mother's educational level were as follows: high school graduate and above, middle class; from ninth to eleventh grades, lower class; and grades 8 and below, lower-lower class. Meeting at least two out of the three criteria mentioned above were necessary for a S to be classified as being in a given social class.

INSTRUMENTS

The instruments used in this investigation were a questionnaire which collected data concerning the S's age, race, grade level, father's and mother's educational levels, and father's occupation; Haller's Occupational Aspiration Scale (OAS); the NORC Occupational Prestige Scale (OPS); the McClelland's (6) Need Achievement Thematic Apperception Test (TAT); and the Mandler-Cowan's Test Anxiety Questionnaire (TAQ) for High School Students.

The OAS (4) is an 8-item multiple-choice instrument designed primarily for use among male high school students. The total score is interpreted as a relative indicator of the prestige level of the occupational hierarchy which an individual views as a goal. The reliability of this scale is reported to be about .80.

The OPS consisted of twenty occupations which were selected from the list of ninety used in the NORC study (7). Subjects were instructed to rank these twenty occupations on the basis of their opinion as to which occupation had the most prestige. Scoring was done by subtracting the ideal rank provided by the scale from the rank given by the S, or vice versa, for each occupation. The total score was then obtained by adding the discrepancy scores for all the twenty occupations. This represented the deviation of the S's perception of occupational prestige in relation to the social norm.

The TAT consisted of four pictures (1) presented in a neutral classroom situation. Scoring was done by two trained graduate students with inter-rater reliability of .90.

A short form of TAQ (5) consisted of thirty-two items. It correlated .946 with the 48-item form. Each item was graded on a 9-point scale with 1 representing low anxiety level and 9 representing high anxiety level.

RESULTS

With social class (middle, lower, and lower-lower) and grade level (9, 10, 11, and 12) as two factors, analyses of variance on a 3x4 factorial design were carried out with regard to the four dependent measures—occupational aspiration, perception of occupational prestige, achievement motivation, and fear of failure. Results of these analyses are shown in Table 1.

Social class as an independent variable was found to be the only significant main effect (p < .001) for both occupational aspiration and perception of occupational prestige. A significant interaction effect (p < .05) was found on achievement motivation, but no significant main effects or interaction were found on fear of failure.

The significant interaction found on achievement motivation indicated that a given social class had different effects for one grade level of Ss from what it had for other grade levels and that a given grade level had different effects for one social class of Ss from what it had for other social classes. In order to test the simple effects of the social classes for each of the grade levels and those of the grade levels for each of the social classes, variance analyses were carried out. [3] Table 2 summarizes the results.

It was found that there were significant differences between social classes for grade 9 and that there were significant differences between grade levels for the lower-lower class, as far as achievement motivation was concerned.

To further examine the mean differences of the three socioeconomic groups (M, L, LL) on occupational aspiration and perception of occupational prestige as well as the mean differences of the three social classes of the ninth grade subjects (9M, 9L, 9LL) and the four grade levels of the lower-lower class Ss (9LL, 10LL, 11LL, 12LL) on achievement motivation, the Duncan's New Multiple Range Test was used. Results of these multiple comparisons are given in Table 3.

The mean occupational aspiration score of the M group was significantly higher (p < .05) than those of the L and LL groups. There was no significant mean difference between the L and LL groups on occupational aspiration. In other words, Ss from the lower socioeconomic classes showed significantly lower occupational aspiration than those from the middle class.

The mean score of the perception of occupational prestige of group M was significantly (p < .05) lower

TABLE 1

ANALYSES OF VARIANCE OF FOUR DEPENDENT VARIABLES

Variable and Source	df	MS	F
Occupational aspiration (OAS)			
Social class	2	2029.21	16.62*
Grade level	3	214.18	1.75
S x G	6	191.82	1.57
Residual	167	122.10	
Occupational prestige (OPS)			
Social class	2	12006.65	9.11*
Grade level	3	889.55	0.68
S x G	6	1278.11	0.97
Residual	167	1318.66	
Achievement motivation (TAT)			
Social class	2	36.99	1.12
Grade level	3	59.71	1.81
S x G	6	70.72	2.15**
Residual	166	32.96	
Fear of failure (TAQ)			
Social class	2	2760.23	2.54
Grade level	3	235.89	0.22
S x G	6	808.19	0.74
Residual	164	1089.05	

*$p < .001$
**$p < .05$

than those of groups L and LL which were found to be homogeneous. It appeared that as the social class of the group shifted away from the middle class toward the lower socioeconomic levels, the Ss' perception of occupational prestige hierarchy became significantly more distorted from a national norm (7) standpoint.

Significant mean differences ($p < .05$) found in achievement motivation are as follows. Among the ninth grade Ss, 9M possessed significantly higher achievement motivation than the 9L and 9LL groups.

TABLE 2

ANALYSES OF VARIANCE OF ACHIEVEMENT MOTIVATION FOR SIMPLE EFFECTS

Source	df	MS	F
Social class for grade 9	2	151.50	4.59*
Social class for grade 10	2	36.50	1.11
Social class for grade 11	2	14.50	0.44
Social class for grade 12	2	34.00	1.03
Residual	166	32.96	
Grade for middle class	3	21.67	0.66
Grade for lower class	3	74.67	2.26
Grade for lower-lower class	3	104.67	3.18*
Residual	166	32.96	

*$p < .05$

TABLE 3

MULTIPLE COMPARISONS AMONG MEANS WITH $p = .05$

Variable	Mean Comparison*		
Occupational aspiration	L	LL	M
	37.1 (n=64)	39.0 (n=57)	48.0 (n=58)
Occupational prestige	M	L	LL
	66.7 (n=58)	90.8 (n=64)	92.0 (n=57)
Achievement motivation	9LL	9L	9M
	5.7 (n=6)	7.4 (n=9)	13.1 (n=14)
Achievement motivation	9LL	10LL	11LL	12LL
	5.7 (n=6)	8.8 (n=25)	10.1 (n=9)	13.1 (n=17)

*Groups underlined by the same line are homogeneous, and any pair of group means not underlined by the same line are significantly different at the .05 level.

Whereas, among the Ss from LL social class, the twelfth grade Ss showed significantly higher achievement motivation than the 9LL and 10LL Ss.

In order to determine the extent to which each of the four dependent variables might be related to the others, correlational analyses were carried out. The resultant product-moment correlation coefficients are shown in Table 4.

The only nonsignificant correlation coefficient found was that (-.10) between the scores of achievement motivation (TAT) and fear of failure (TAQ). This finding confirms Atkinson and Litwin's (2) report that the measure of achievement motivation obtained from TAT and the measure of fear of failure obtained from TAQ are uncorrelated.

TABLE 4

INTERCORRELATIONS (N=176)

Variable	Occupational aspiration	Occupational prestige	TAT	TAQ
Occupational aspiration	-			
Occupational prestige	-.50***	-		
TAT	.22*	-.26**	-	
TAQ	-.26**	.25**	-.10	-

*$p < .05$
**$p < .01$
***$p < .001$

TABLE 5

ANALYSES OF COVARIANCE

Variable and Source	df	MS	F
Occupational aspiration (criterion)			
Occupational prestige (covariance)			
Social class	2	950.23	8.88*
Grade level	3	320.38	2.39
S x G	6	124.13	1.16
Regression	1	3801.78	
Residual	166	107.02	
Occupational prestige (criterion)			
Occupational aspiration (covariance)			
Social class	2	2306.31	2.11
Grade level	3	1757.75	1.61
S x G	6	930.10	0.85
Regression	·1	38819.77	
Residual	166	1092.75	

*p<.001

Occupational aspiration and distortion of perception of occupational prestige hierarchy were found to have a rather high and negative correlation ($r=-.50$, $p<.001$). Since occupational aspiration and perception of occupational prestige seemed to covary, analyses of covariance on a 3x4 factorial design were conducted with regard to each of the two variables using the other as the covariate. Table 5 presents the results.

When occupational aspiration was adjusted in terms of the differences in the Ss' perception of occupational prestige hierarchy, exactly the same findings as revealed by the earlier variance analysis resulted. In other words, social class was the only significant main effect and Ss from the lower and lower-lower socioeconomic groups showed significantly lower occupational aspiration than Ss from the middle class, with differences in their perception of occupational prestige being controlled.

When the Ss' perception of occupational prestige was adjusted in accordance with the differences in their occupational aspiration, none of the main effects and interaction was found to be statistically significant.

SUMMARY AND DISCUSSION

The sample of this study consisted of 179 male students of grades 9 through 12 from a geographically isolated and culturally deprived area which is located in what is commonly known as the Appalachian Poverty Belt. This research has attempted to answer questions concerning the relationship of socioeconomic levels and grade levels to the variables occupational aspiration, perception of occupational prestige, achievement motivation, and fear of failure.

Variance analyses, covariance analyses, and multiple comparisons revealed that socioeconomic groups differed significantly on occupational aspiration, with or without their differences in perception of occupational prestige being controlled, that socioeconomic groups differed significantly on perception of occupational prestige hierarchy without controlling for their differences in occupational aspiration, that socioeconomic levels of the ninth grade group differed significantly on achievement motivation, that grade levels of the lower-lower socioeconomic group differed significantly on achievement motivation, and that there were no significant differences between socioeconomic groups or between grade levels on fear of failure.

It appears that the individual's socioeconomic status does have a great deal to do with occupational aspiration, perception of occupational prestige, and achievement motivation. In order to stimulate social mobility in the lower socioeconomic groups, therefore, it would be necessary to help improve the goal they set for the attainment of higher level occupations, change their perception concerning the world in general and on the world of work and education in specific, and acquire a stronger urge to approach success.

Of an empirical interest was the degree to which the four dependent variables investigated in the study, occupational aspiration, perception of occupational prestige, achievement motivation, and fear of failure, might be tapping the same factors. A factor analytic approach was, thus, used to clarify this point. The principal-component solution was used to generate a factor matrix from the 4x4 correlation matrix. The extracted factors I, II, III, and IV contributed 46, 22, 20, and 12 percent of the total variance, respectively. These factors were then orthogonally rotated to optimize the contribution of each of the four variables to each of the four factors. Results are shown in Table 6.

Factor I is characterized by a high factor loading of .96. This high correlation between Factor I and occupational aspiration together with other insignificant factor loadings of -.25, .09, and -.11 would indicate that this is the occupational aspiration factor. Factor II, represented by a high factor loading of .99, clearly is the achievement motivation factor. It can be observed from Table 6 that, in fact, Factor III is the fear of failure factor and Factor IV is the perception of occupational prestige factor. In conclusion, the four dependent variables tapped by this study in relation to socioeconomic and grade levels of adolescent boys appear to be uniquely different variables.

TABLE 6

ROTATED FACTOR MATRIX

Variable	Factor I	Factor II	Factor III	Factor IV
Occupational aspiration	.96	.10	-.12	-.25
Occupational prestige	-.25	-.12	.12	.95
Achievement motivation	.09	.99	-.04	-.11
Fear of Failure	-.11	-.04	.99	.11

FOOTNOTES

1. This study was supported in part by the Office of Economic Opportunity as a part of the Mc-Dowell County Evaluation Project, Contract Number OEO-703 between West Virginia University and the Office of Economic Opportunity.

2. The author would like to express his appreciation to Donald L. Thompson for his assistance.

3. A method concerning analysis of variance for simple effects of a statistically significant interaction can be found, for example, in B. J. Winer's Statistical Principles in Experimental Design, McGraw-Hill, New York, pp. 233-238.

REFERENCES

1. Atkinson, J. W., Motives in Fantasy, Action, and Society, D. Van Nostrand Company, Inc., Princeton, New Jersey, 1958.

2. Atkinson, J. W.; Litwin, G. H., "Achievement Motive and Test Anxiety Conceived as Motive to Approach Success and Motive to Avoid Failure," Journal of Abnormal and Social Psychology, 60: 52-63, 1960.

3. Beilin, Harry, "The Utilization of High Level Talent in Lower Socioeconomic Groups," Personnel and Guidance Journal, No. 35, 1956.

4. Haller, A. O.; Miller, I. W., The Occupational Aspiration Scale Theory, Structure and Correlates, Department of Rural Sociology, The University of Wisconsin, Madison, 1967.

5. Mandler, G.; Cowen, J., "Test Anxiety Questionnaires," Journal of Consulting Psychology, 22:228-229, 1958.

6. McClelland, D. C., The Achieving Society, D. Van Nostrand Company, Inc., Princeton, New Jersey, 1961.

7. National Opinion Research Center, "Jobs and Occupations: A Popular Evaluation," Opinion News, 9:3-13, 1947.

8. Tseng, M. S.; Thompson, D. L., "Differences Between Adolescents Who Seek Counseling and Those Who Do Not," The Personnel and Guidance Journal, 47:333-336, 1968.

9. Tseng, M. S.; Thompson, D. L., "Quarterly Report: The Occupational Study," McDowell County Evaluation Project, West Virginia University, Morgantown, 1968.

THE JOURNAL OF EXPERIMENTAL EDUCATION

A QUARTERLY

DIRECTIONS FOR J.E.E. CONTRIBUTORS

The Journal of Experimental Education publishes specialized or technical education studies, treatises about the mathematics or methodology of behavioral research, and monographs of major current research interest.

ABSTRACT REQUIRED

An abstract of not more than 120 words must accompany each manuscript. It should precede the text, be designated *ABSTRACT,* and conform to the following standards:

1. In a research paper, include statements of (a) the problem, (b) the method, (c) the data, and (d) the conclusions.

2. In a review or discussion article, state the topics covered and the central thesis.

3. Standard abbreviations may be used freely if the meaning is clear. Use complete sentences and do not repeat information which is in the title.

TEXT

Usually research articles should have a clearly organized order of presentation. The following elements and their order is a guide and is not intended to be prescriptive.

The Problem. The nature, scope, and significance of the problem should be presented.

Related Research. Presentation and discussion of related research should be selective and should include references essential to clarify the problem under examination.

Methodology. This section should consist of hypotheses, description of the sample and sampling procedures, discussion of the variables, description of the data-gathering instruments (if well-known, their description may be omitted), and general description of the statistical procedures.

Presentation and Analysis of Data. Analysis of the data and conclusions about the hypotheses should be more than presentation. Rival hypotheses and significance for related studies and educational theory should be considered. Summary data (means, standard deviations, frequencies, correlation coefficients, etc.) should be presented in tabular or graphic form. Discussion of these data should be interpretive.

Summarizing Statements. A summary of conclusions and implications for education may supplement the abstract.

STYLE

Certain suggestions are offered here to achieve unity in publication style and to conserve the time and energy of the author and editorial staff in the preparation of copy. *A Manual on Writing Research,* 1962, and *Manual of Form for Theses and Term Reports,* 1962, by Kathleen Dugdale, the Indiana University Bookstore, Bloomington, may be used as style manuals in preparation of manuscripts.

Two Copies Required. Copies should be double spaced with wide margins. Typed copies are preferred, but dittoed or mimeographed copies will be accepted if they are legible.

Subheads. Articles are frequently improved by the judicious use of subheads. However, avoid use of the title, INTRODUCTION, for a lead section.

Title. Try to use a short title, preferably no more than 7 words. Avoid superfluous phrases, such as "A Comparison of . . .," "A Study of . . .," and "The Effectiveness of . . . ."

Tables. Prepare tables precisely as they are to appear in *The Journal,* caption each with a brief and to-the-point title, and number consecutively with arabic numerals: *Table 3.* INTERCORRELATION MATRIX, VARIABLES OPTIMALLY ORDERED.

Figures. Draw any graphs and charts in black ink on good quality paper, title each, and number consecutively, thus: *Figure 4.* SCHOOL ENROLLMENT. Send the original copy. We cannot accept mimeographed figures or charts. Data should not be framed, and ordinates and abscissas should be shortened to occupy as little space as possible.

Tables and Figures. Tables and figures must be original copies acceptable for reproduction. A charge will be assessed for any redrawing or re-typing of tables or figures.

Technical Symbols. All symbols, equations, and formulas must be clearly represented. Differentiate between numbers and letters (zero and the letter o; one and the letter l, etc.) Identify hand-drawn Greek letters in the margin. Align subscripts carefully.

Footnotes. Avoid explanatory footnotes by incorporating their content in the text. However, for essential footnotes, identify them with consecutive superscripts, as *book,*[2] *study,*[3] etc., and list the footnotes in a section, entitled FOOTNOTES, at the end of the text, but preceding the REFERENCES.

References. References should be listed alphabetically according to the author's last name at the end of the manuscript. Each entry should be numbered. Citation of a reference in the text should be made with this number, as (2); or if specific pages are to be indicated, as (2:245–7).

Illustrations of article and book references:

1. Bittner, Reign H. and Wilder, Carlton E., "Expectancy Tables: A Method of Interpreting Correlation Coefficients," *The Journal of Experimental Education,* 14:245–52, March 1946.
2. Thomson, Godfrey, H., *The Factorial Analysis of Human Ability,* Houghton Mifflin Co., Boston, 1950, 383 pp.

COSTS

The publisher charges a contributor's fee of $6 per printed page of approximately 1,200 words, billed upon publication. Authors are charged for changes in tables, figures, or copy made when article is in camera-ready form. Each contributor will receive 10 complimentary copies of the issue in which his article appears. Reprints are charged at cost, and a price schedule will be sent to each contributor.

PROOFREADING

We will send you proofs for correction (with instructions for handling). Any major changes made in the proofs that were not incorporated in your original copy will be an added expense to you. (Errors that we make, naturally, will be at our expense.)

Keep an exact carbon copy of your manuscript so that if a question arises the editors can refer you to specific pages, paragraphs, or lines for clarification by letter.

SEND MANUSCRIPTS TO

John Schmid, Department of Research and Statistical Methodology, University of Northern Colorado, Greeley, Colorado 80631.

THE *Journal* OF
Experimental
Education

CONTENTS

CONCEPT SELECTION STRATEGIES OF

NEW GUINEA STUDENTS

A. L. ALLEN
University of Papua and New Guinea

A. G. SHANNON
New South Wales Institute of Technology

ABSTRACT

Some groups of preliminary year indigenous students at the University of Papua and New Guinea were Ss for an investigation of concept selection strategies among students from a non-Western culture. Some of the methods of Bruner, Goodnow, and Austin (1) were replicated. It was found that in forming conjunctive categories the students were consistent in maintaining a definite strategy, that more students adopted a scanning strategy, and that focusers were the most successful. Mixed strategists attained no success. The abstraction of disjunctive concepts provided more difficulties, as it did with their Western counterparts. The Ss, especially focusers, did not maintain their strategies. Focusers and scanners on the conjunctive problems did not interchange roles on the disjunctive problems, but some of each adopted mixed strategies with the latter problems. Mixed strategists and focusers were more successful than scanners, though focusers either solved all or no problems. The Ss generally found the use of negative examples difficult. The investigation is the first part of a series of studies.

THE RESULTS described in this paper are the initial ones from the first part of a longitudinal study. It is felt that the results so far obtained are of sufficient interest to warrant discussion now. The objectives of these experiments are to investigate the concept selection strategies of indigenous mathematics students at the University of Papua and New Guinea, to describe these strategies mathematically, to link this mathematical formulation with a category model of abstraction, and to contrast the results of Australian students with Papuan and New Guinean students. The experiments were also carried out to provide background information for a mathematics project.

Bruner, Goodnow, and Austin (1) made reference to, the difficulty that the Harvard students encountered with disjunctive categories. They suggested that this may be a legacy of their Western culture with its "common effects have common causes" style of thinking, as well as a general inability to use negative information. In this paper, some initial findings about strategies employed by Papuan and New Guinean students are presented as a source of some information about a less-Westernized culture.

RELATED RESEARCH

Bruner and others have carried out the classic study of idealized concept selection strategies. Their Ss had to form conjunctive, disjunctive, and relational categories when presented with positive and negative examples. In their terms, attaining a concept and learning a category were identical. The two ideal strategies they describe were called focusing and scanning.

The pioneering study in the comparison of performances in a conceptual task with some ideal strategy was that of Whitfield (9). He was followed by Hovland and Weiss (6), and like Bruner and others, found that negative information was not used as efficiently as positive information in concept attainment.

A criticism by Henle (5) of Bruner and others was that they tended to under-emphasize the role of logic in their investigations of reasoning processes. Previously, Donaldson (3) had also criticized their failure to discuss perceptual links in the use of negative information. Bruner and others suggested that to avoid mistakes students distrust transformation of negative to positive information.

Donaldson's research involved repeated transformations in both directions. His Ss found the sequence negative to positive and back especially difficult. The difficulty was in using the positive derived from the negative.

His explanation for this was twofold: the air of finality of positive statements and a certain mistrust of negative information. This latter point was accounted for by Donaldson not in terms of the risk of failure so much as a less rational feeling "that negative information is not such good currency as positive information."

Campbell (2) developed the work of Bruner and others and Donaldson and supported another suggestion of the former, that lack of facility with indirect procedures may extend to a variety of cognitive activities other than categorization.

All of these studies have tended to confuse two variables: the learning of rules when one knows the relevant properties in a concept, and the learning of the relevance of the properties in a concept when one knows the rules. Haygood and Bourne (4) showed that there was a difference between these two variables. Wallace (8) summarizes Bruner's work and mentions two other studies that support it.

THE PRESENT STUDY

Attempts were made in this study to replicate certain aspects of Bruner's work at Harvard. The two types of category considered here were the conjunctive and disjunctive. Eight examples of each concept to be attained were presented to Ss. These examples were either positive (some aspect of the concept) or negative (no aspect of the concept). Each example confirmed or contradicted the S's hypothesis.

This hypothesis was formed by considering all or some of the attributes of the first positive example. Those who considered all the attributes were using a focusing strategy. Those who considered only one attribute at a time were using a scanning strategy. The others were adopting a mixed strategy. The two ideal strategies in forming conjunctive categories are schematically described in Figure 1. In the example in Figure 1, if the concept to be attained were conjunctive it would be "things containing c and d"; if it were disjunctive it would be "things containing c, or things containing d, or things containing c and d."

In abstracting the conjunctive concept, the S takes the common attributes between successive positive examples. Each negative example can be replaced by a complementary positive example, of which the

FIGURE 1

IDEAL STRATEGIES FOR CONJUNCTIVE CATE-
GORIES

a, b, c, d are attributes of the first positive example. Successive positive examples confirm or contradict a, b, c, d.
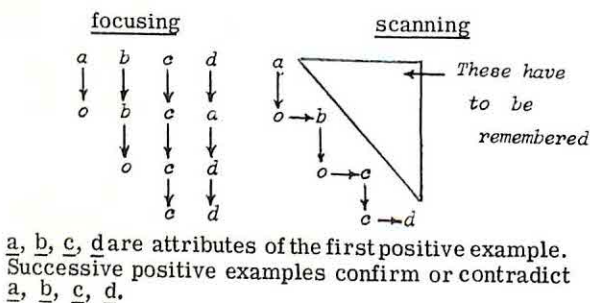
FIGURE 2

DESCRIPTION OF CONJUNCTIVE AND DISJUNCTIVE CATEGORIZATION

A, B, C, D are positive examples;
E, F, G, H are negative examples.

E implies the existence of a complementary positive cE, consisting of those attributes which are not in E.

Conjunctive category is formed from:
(A or B or C or D) and (cE or cF or cG or cH).

Disjunctive category is formed from:
(A and B and C and D) or c (E and F and G and H).

---

attributes are those that are not in the negative example. These are used as described in Figure 2. Figure 2 also shows the formation of the disjunctive category where use is made of the complementary set formed from all the attributes in the negative examples. The descriptions in Figure 2 are based on a set theoretic analysis of categorization and are not unique. For instance, c (E and G and H and I) can be replaced by (cE or cG or cH or cI).

Trial runs were made with different forms of the experiment on some Australian undergraduate mathematics students and then on some Papuan and New Guinean meteorology students. In the form that was adopted, the testing was done in a group situation with the cards of Bruner and others projected onto a screen. Each card was projected for 20 seconds, whereas Bruner and others had displayed them for 10 seconds. It was felt that some allowance had to be made for the fact that English was the second language for many of the Ss.

These cards could have one, two, or three borders, with one, two, or three shapes which could be circles, crosses, or squares and in one of three colors. There were eighty-one cards which were tested in successive weeks.

The students' responses were recorded on specially prepared record sheets. The group situation required the use of several monitors to assist in the administration of the test. Very few difficulties were encountered that had not been foreseen as a result of the trial runs.

THE POPULATION

The University of Papua and New Guinea has a preliminary year which is a year between the completion of the Papua and New Guinea School Certificate and the commencement of undergraduate studies. It was felt that the preliminary year students would be the least Westernized.

The preliminary year classes had been matched by the University administration and three were chosen at random for testing. There is no reason to believe that they were not typical preliminary

TABLE 1

CONJUNCTIONS

Strategy	Scanning	Focusing	Mixed
Number $X^2 = 17.9$ $p < 0.001$	39 (60%)	22 (35%)	4 (5%)
4/4 correct	8 (21%)	16 (73%)	0 (0%)
Secondary Schooling: Administration	20 (31%)	10 (15%)	1 (3%)
Mission	19 (29%)	11 (17%)	3 (5%)
Mean Age	18 years 8 months	17 years 11 months	18 years 6 months

year students, especially because Kearney (7) points out that they are more homogeneous than their Western counterparts.

THE RESULTS

Table 1 shows the results for the conjunctive concepts. The first feature to be noticed is the high percentage (95%) of students who adopt one strategy or the other consistently in forming the conjunctive category. More go in for scanning than focusing in contrast to the Harvard group. The next point that emerges is the greater success of the focusers in solving the problems. They are also younger on the average and there are no focusers over the age of 19, though it is not possible to infer anything from this evidence. If we try to determine whether the expected frequencies are equal for all categories we have to accept H_1 (the frequencies are not equal) and reject H_0 (the frequencies are equal and X^2 is distributed as chi-square with 2 d.f., $\alpha = .05$).

Some attempt was made to see if there was any relation between the strategy employed and the district in which the student lived as a child but no

TABLE 2

CONJUNCTIVE STRATEGIES COMPARED WITH DISTRICT OF PRIMARY SCHOOL EDUCATION AND WITH SUBJECTS' PROPOSED UNDERGRADUATE COURSES

Strategy	Scanning	Focusing	Mixed	X^2
Papua New Guinea Mainland	7	8	1	5.8
New Guinea Mainland	10	6	1	6.8
New Guinea Islands	22	8	2	21.2*
Arts-Law- Education	12	8	2	7.3
Science- Dentistry- Medicine	27	14	2	22.4*

*p < 0.001

TABLE 3

STRATEGIES FOR DISJUNCTIVE CATEGORIES

Strategy	Scanning	Focusing	Mixed
number (%) $X^2 = 9.9$ $p < 0.01$	25 (38%)	10 (16%)	30 (46%)
4/4 correct	0 (0%)	5 (50%)	5 (17%)
0/4 correct	10 (40%)	5 (50%)	20 (67%)

definite picture emerged. A crude summary is provided in Table 2. The main feature of interest in this part was that all the East New Britain students were scanners. It is intended to investigate whether this is related to the fact that in a recent trial of mathematics learning materials for grade 7 in Papua and New Guinea, the only students to encounter serious difficulties in set theory and logic were those at schools in East New Britain. Little can be learned at this stage from the rest of Table 2 which compares the strategies and the proposed undergraduate courses of the Ss. Chi-square tests for individual classifications are also displayed in Table 2. Cross-classifications gave $X^2 = 4.7$ (4 d.f.) and $X^2 = 2.7$ (2 d.f.) for the districts and courses respectively, neither of which were significant.

Table 3 shows the results on the disjunctive problems. Like their Western counterparts, the Ss found much more difficulty with these. The chi-square test showed a significant difference.

Many of the focusers and scanners of the conjunctive problems adopted mixed strategies when confronted with the use of the negative examples in the disjunctive problems. Focusers on the conjunctive problems did not become scanners, or vice versa. Focusers tended to get all or none of the problems correct, and were, with the mixed strategists, more successful than the scanners. Conjunctive focusers were no more successful as disjunctive mixed strategists than conjunctive scanners.

CONCLUDING DISCUSSION

It should be borne in mind that the Ss were not untouched by Western culture. Although the materials of the experiment were virtually culture-free, it would not be easy to carry out similar testing on villagers in remote areas.

No explanation can be offered yet in cultural terms for the remarkably high percentage of adherents to a definite strategy on the conjunctive items, or the relative success of the focusers, or the preference for scanning rather than focusing. In other respects the Ss did not seem to differ much from their Western counterparts, though later testing of the same Ss and more refined techniques for testing new Ss may enlarge any differences.

REFERENCES

1. Bruner, J. S.; Goodnow, J. J.; Austin, A. G., A Study of Thinking, Wiley, New York, 1956.

2. Campbell, A.C., "On the Solving of Code Items Demanding the Use of Indirect Procedures," British Journal of Psychology, 56:45-51, 1965.

3. Donaldson, M., "Positive and Negative Information in Matching Problems," British Journal of Psychology, 50:235-262, 1959.

4. Haygood, R.C.; Bourne, L.E., Jr., "Attribute- and Rule-learning Aspects of Conceptional Behaviour," Psychological Review, 72:175-195, 1965.

5. Henle, M., "On the Relation between Logic and Thinking," Psychological Review, 69:366-378, 1962.

6. Hovland, C.I.; Weiss, W., "Transmission of Information Concerning Concepts through Positive and Negative Instances," Journal of Experimental Psychology, 45:175-182, 1953.

7. Kearney, J.E., "Success Factors in a Tertiary Institution," Australian Journal of Higher Education, 3:231-236, 1969.

8. Wallace, J.G., Concept Growth and the Education of the Child, National Foundation for Educational Research in England and Wales, 1965, pp. 27-30.

9. Whitfield, J.W., "An Experiment in Problem Solving," Quarterly Journal of Experimental Psychology, 3:184-197, 1951.

AN INSTRUMENT FOR ASSESSING TEACHERS'

ATTITUDES TOWARD INDIVIDUALIZING

READING INSTRUCTION

EUNICE N. ASKOV
The University of Wisconsin

ABSTRACT

The development of an instrument for measuring teachers' attitudes toward individualizing reading instruction is described. The instrument was constructed in the semantic differential format c h o s e n as an indirect means of assessing attitudes. Teachers were asked to respond anonymously on adjective scales to eleven examples of classroom procedures. The examples were designed to be applications of the assumptions of individualized reading instruction. Reliability of the instrument was high (.93); content validity w a s demonstrated f o r classroom examples and adjectives used in the instrument. Two validation studies which established the effectiveness of the instrument in discriminating among teachers' attitudes toward individualizing reading instruction are also reported.

THE IMPORTANCE of individualizing reading instruction —also known as " diagnostic teaching" or " individually guided instruction "—has been recently emphasized. Regardless of terms, the notion is that instruction should be based on pre-assessment of children's individual strengths and weaknesses. In order for this type of instruction to become a classroom reality, teachers must focus on the needs of individuals rather than the group. This orientation may require a basic change in teachers' attitudes as the need for continuous assessment and modification of each individual's instructional program based on the assessment take precedence over getting the classroom group through the pages in a basal reader and accompanying workbook.

One attempt to assess teachers' predisposition toward various instructional approaches in reading has come to the author's attention. The San Diego Teacher Inventory of Approaches to the Teaching of Reading (8) measures teachers' agreement with the assumptions of three instructional approaches—basic, individualized, and language experience. The definition of the individualized approach, however, is the classic one advocated by Veatch (9) and others, involving the principles of seeking, self-selection, and self-pacing. The San Diego Inventory, consequently, does not measure teachers' attitudes toward the most recent concept of individualization—that of planning each child's reading instructional program based on his pre-assessed needs—which can occur within any instructional approach, with any group size, using any materials.

An instrument was thus constructed to assess teachers' attitudes toward individualizing instruction in reading. In this paper the development of the instrument and two validation studies are described.

DEVELOPMENT OF THE INSTRUMENT

A problem in attitude assessment has been that respondents tend to give the answers they think are expected, rather than to respond as they actually believe. Jackson and Messick (4) cautioned that " indirect, disguised techniques " are sometimes necessary to obtain a valid measurement of an attitude. Weschler and Bernberg (10:225), furthermore, stated that " to a certain extent the value of a given technique may depend upon the manner in which it is able to disguise its true purpose and can be adjusted

to fit into a variety of different situations."

An indirect method of assessing attitudes, the Reading Teacher Survey, was accordingly chosen for the form of the attitude inventory. It was an adaptation of the semantic differential.

Remmers (7), after summarizing several studies that employed the semantic differential in assessing attitudes for various purposes, cautioned that a bias due to response-sets may be operating. In other words, the order of presentation of the concepts to be evaluated may influence the responses of the S. More recently, however, Kane (5), after analyzing data from a semantic differential instrument which included various combinations for ordering items, showed that item order is not a significant influence and that an experimenter need not worry about proximity errors.

Osgood, Suci, and Tannenbaum have discussed the flexibility of the semantic differential:

Although we often refer to the semantic differential as if it were some kind of "test," having some definite set of items and a specific score, this is not the case. To the contrary, it is a very general way of getting at a certain type of information, a highly generalizable technique of measurement which must be adapted to the requirement of each research problem to which it is applied. There are no standard concepts and no standard scales; rather, the concepts and scales used in a particular study depend upon the purposes of the research. (6:76)

Two adaptations of the basic semantic differential instrument, as described by Osgood and others (6), were made here. First, analysis of the three factors found by Osgood and others—evaluation, potency, and activity—was not undertaken since measurement of a unitary concept of attitude toward individualizing reading instruction seemed more desirable. Second, an agree-disagree scale was included to determine whether Ss would tend to respond more positively to it than to the other scales which consisted of adjectives. This notion was supported by the data.

Pilot Studies

The form of the Reading Teacher Survey was changed several times before the final version was arrived at. A brief summary of changes is given here; a more detailed description may be found elsewhere (1).

The first version consisted of twelve statements which summarized the basic tenets of individualizing instruction in reading. Subjects were asked to respond to the statements on seven adjectival scales and agree-disagree scale, each of which had seven positions ranging from the negative extreme to the positive extreme. The adjectives for the scales were picked from those used in the literature to describe individualized reading instruction. This version, however, was not retained after a pilot test because teachers all tended to mark the positive extreme, and,

thus, the instrument did not discriminate among individuals. It appeared that teachers were responding positively to the theories of individualization without thinking of the classroom ramifications.

The next version presented twelve examples of classroom situations, illustrating procedures that would grow out of the assumptions of individualized instruction. One of the twelve statements or examples, representing a viewpoint opposed to individualization, was inserted to break a set toward positive responses (see Appendix, statement 4).

Teachers, who did not sign their names, were asked to consider the feasibility of applying each of the twelve examples in their classrooms. They were instructed to record their responses on the eight rating scales—the same ones used in the previous version—following each example. This version, after pilot testing and subsequent minor revisions, was used in a study of attitude change (1).

Revised Version

Since the instrument did reflect a change in teachers' attitudes due to an intervening experimental treatment, further revisions were made on the basis of item analyses. Using the Generalized Item Analysis Program (2), each of the ninety-six items (i. e., each of eight scales under each of the twelve examples) was analyzed both in terms of the correlation with the subtest score (the subtest being the classroom example) and in terms of the correlation with the total test score.

Extensive revisions were made. The instrument was shortened to sixty-one items or scales. One classroom example was eliminated and one was rewritten, making a final total of eleven. Particular scales that did not discriminate between high and low total scores were also omitted; therefore, only the scales that had the highest correlations with total score were used under each example in the third version. One adjective scale that proved to be ineffective was omitted, leaving a total of seven possible scales.

Respondents were asked to rate, on the scales provided, the feasibility of applying each classroom example. The scales had seven positions ranging from the negative extreme to the positive extreme; the middle position could be used when the respondent felt neutral or did not know how to answer. With the exception of asking teachers to consider each statement in terms of their classroom experience, the instructions were modeled after those suggested by Osgood and others (6:82-84).

By summing the point values or responses on each scale, the test yielded a total possible score of 441 points. (The most positive response was awarded 7 points and the most negative response was scored as 1 point; the five positions between the extremes were awarded 2 to 6 points.) The instrument was then given to thirty-one experienced elementary school teachers; the data were again submitted to item analysis, but only very minor revision were made. The statements and scales that were used in the Revised Version are presented in the Appendix.

Validity and Reliability

Content validity was demonstrated for the classroom examples and adjective scales. Three professors who teach courses in reading at the University of Wisconsin judged the classroom examples to be relevant to measuring attitudes toward individualizing reading instruction. They, furthermore, judged that the classroom examples represented application of three assumptions of individualized instruction:

1. The teacher should do diagnostic testing to determine the specific strengths and weaknesses of each child.

2. Children should receive instruction in the skills they need and move at their own pace.

3. Children should be given materials appropriate to their abilities and interests.

The same professors also judged the scales to be relevant to measuring attitudes toward individualizing reading instruction. The scales had further validity in that the adjectives were chosen from the literature describing individualized reading instruction.

The estimate of reliability or internal consistency (Hoyt reliability coefficient) of the Revised Version was . 93 based on the data gathered from the thirty-one experienced elementary school teachers. Green (3) has stated that a high reliability coefficient usually indicates that the items are homogeneous and the scales unidimensional. Since the reliability coefficient for the Revised Version was high and since, therefore, the items were highly intercorrelated, the instrument was apparently measuring one factor as was intended instead of the three factors found by Osgood and others (6). Presumably, this unitary factor was teachers' attitudes toward individualizing reading instruction.

VALIDATION STUDIES

Two types of questions were asked in seeking experimental validation of the instrument: (1) Could the instrument successfully discriminate between the attitudes of teachers who were systematically individualizing reading instruction and those of teachers who were not? (2) Could it measure a change in teachers' attitudes when instruction in a school had changed from conventional to individualized?

Study 1

The instrument was given to the teachers in two types of schools. The first type (Type 1) had successfully implemented an experimental system for individualizing reading instruction at least a year prior to the study. Teachers in Type 1 schools were systematically assessing pupil needs and planning instruction accordingly. In the second type of school (Type 2) there had been no known emphasis on individualizing reading instruction. The hypothesis was that scores on the attitude inventory would be significantly higher in Type 1 schools than in Type 2 schools.

The Reading Teacher Survey, Revised Version,

was administered in the fall of 1969 in two Type 1 schools and five Type 2 schools in small or middle sized Wisconsin cities. All classroom teachers of grades 1-6 took the inventory; special teachers—such as reading teachers—were not included.

Analyses and Results. A t-test was performed on the data, using an estimate of the variance of the means. The following formula was used for pooling the variance of the means of Type 1 schools with that of Type 2 schools:

$$s^2 = \frac{(\bar{x}_1 - \bar{x}_2)^2}{2} + \Sigma (\bar{y} - \bar{y}_i)^2$$
$$\overline{}$$
$$v$$

$\bar{x}_1$ = mean of scores in Type 1 School No. 1

$\bar{x}_2$ = mean of scores in Type 1 School No. 2

$\bar{y}$ = grand mean of scores in Type 2 schools

$\bar{y}_i$ = means of scores in Type 2 schools

v = degrees of freedom (number of schools minus 2)

The . 05 level for a two-tailed t-test was designated as the level of significance for testing the difference between means. The following formula was used in figuring the t value:

$$t = \frac{\bar{x} - \bar{y}}{s}$$

Table 1 presents the means and standard deviations of scores for teachers in the seven schools.

TABLE 1

MEANS AND STANDARD DEVIATIONS FOR INVENTORY SCORES IN TYPE 1 AND 2 SCHOOLS

Schools	Mean	Standard Deviation	N
Type 1 Schools:			
No. 1 ($\bar{x}_1$)	388.67	41. 13	21
No. 2 ($\bar{x}_2$)	365.38	24. 15	8
Grand Mean ($\bar{x}$) 377. 02			
Type 2 Schools:			
No. 1 ($\bar{y}_1$)	362. 33	48. 32	21
No. 2 ($\bar{y}_2$)	338. 10	32. 42	21
No. 3 ($\bar{y}_3$)	331. 83	33. 29	6
No. 4 ($\bar{y}_4$)	328. 20	41. 49	15
No. 5 ($\bar{y}_5$)	327. 50	48. 07	10
Grand Mean ($\bar{y}$) 337. 59			

The obtained t value (t = 2.65) from testing the differences between means of inventory scores in Type 1 and Type 2 schools was significant at the .05 level for a two-tailed test. It can also be noted from Table 1 that the ranges in scores of the two types of schools did not overlap.

Study 2

School No. 2 of the Type 2 schools adopted a system for individualizing reading instruction during the 1969-70 school year. Since the Reading Teacher Survey, Revised Version, had been administered the fall before in-service training on individualization was given, the inventory was readministered at the end of the school year to determine if a change in attitudes had occurred after teachers had been systematically individualizing reading instruction for one year. As in Study 1, the instrument was taken by all classroom teachers in grades 1-6; special teachers were not included.

Analyses and Results. Although teachers did not sign their names the inventories taken by each in the fall and spring were paired together by coding the inventories. The means and standard deviations of inventory scores at each administration time are presented in Table 2. (The data in Table 2 for the fall administration are slightly different from the figures given in Table 1 for the same school; several teachers were omitted from the sample in Study 2 because they resigned during the school year.)

A t-test for matched pairs was performed on the data. The obtained t value (t = 4.09) was significant beyond the .001 level for a two-tailed test.

Discussion

Of the two sorts of evidence obtained, the first appears to offer conclusive evidence of validity. Scores on the Reading Teacher Survey, Revised Version, were significantly higher in schools where instruction proceeded from the pre-assessed needs of individual pupils than in schools where individualization was not systematically practiced. The second sort of evidence, scores obtained by the same teachers being significantly higher after they had adopted a system for individualizing reading instruction, is also supportive if one may assume that control groups would not register similar gains.

CONCLUSIONS AND IMPLICATIONS

The Reading Teacher Survey, Revised Version, was shown in two studies to be effective in discrim-

inating among teachers' attitudes toward individualizing reading instruction. Furthermore, the instrument was demonstrated to have a high reliability and content validity. It is easily scored, with a high total score indicating a positive attitude toward individualizing reading instruction. It is quickly administered, taking no longer than 20 minutes; it may be administered to a group with the directions being read aloud, or it may be given to teachers to fill out independently.

The instrument may be used in a variety of situations. It would be especially useful as an evaluative tool in studies of individualized approaches to reading instruction. It may also be used to evaluate the effects of training or some other treatment on the attitudes of teachers. Frequently, teacher opinions and attitudes are solicited only by informal questionnaires; the Reading Teacher Survey, Revised Version, provides a more objective means of assessing such attitudes.

FOOTNOTE

1. The material reported herein was prepared with the support of the Wisconsin Research and Development Center for Cognitive Learning, supported in part as a research and development center by funds from the United States Office of Education, Department of Health, Education, and Welfare. The opinions expressed here in do not necessarily reflect the position or policy of the Office of Education and no official endorsement by the Office of Education should be inferred. Center No. C-03/ Contract OE 5-10-154.

REFERENCES

1. Askov, Eunice N., "Assessment of a System for Individualizing Reading Instruction," Technical Report No. 117, Wisconsin Research and Development Center for Cognitive Learning, The University of Wisconsin, Madison, 1970.

2. Baker, F. B.; Martin, T. J., FORTAP; A Fortran Test Analysis Package, Laboratory of Experimental Design, Wisconsin Research and Development Center for Cognitive Learning, University of Wisconsin, Madison 1968.

3. Green, B. F., "Attitude Measurement," in Jackson, D. N.; Messick, S. (eds.) Problems in Human Assessment, McGraw-Hill, New York, 1967, pp. 725-736.

4. Jackson, D. N.; Messick, S., "Assessment of Attitudes," in Jackson, D. N.; Messick, S. (eds.) Problems in Human Assessment, McGraw-Hill, New York, 1967, pp. 717-723.

5. Kane, R. B., "Minimizing Proximity Errors in the Semantic Differential," paper presented at the 53rd Annual Meeting of the American Educational Research Association, February 1969, Los Angeles, California.

6. Osgood, C. E.; Suci, G. J.; Tannenbaum, P. H.,

TABLE 2

MEANS AND STANDARD DEVIATIONS OF INVENTORY SCORES IN TYPE 2 SCHOOL NO. 2

Administration Time	Mean	Standard Deviation	N
Fall	336.35	32.85	17
Spring	361.12	38.22	17

The Measurement of Meaning, University of Illinois Press, Urbana, 1957.

7.   Remmers, H. H., " Rating Methods in Research on Teaching," in Gage, N. L. (ed.) Handbook of Research on Teaching, Rand McNally, Chicago, Illinois, 1963, pp. 329-378.

8.   Teacher Inventory of Approaches to the Teaching of Reading, San Diego County, Department of Education, Reading Study Project Committee, Monograph No. 4, San Diego California, 1961.

9.   Veatch, Jeanette, Individualizing Your Reading Program, G. P. Putnam's Sons, New York, 1959.

10.   Weschler, I. R.; Bernberg, R. E., " Indirect Methods of Attitude Measurement," International Journal of Opinion and Attitude Research, 4:209-228, 1950.

APPENDIX

Statements and Scales in the Reading Teacher Survey. Revised Version

EXAMPLE:

A. Lucy, Larry, Joe and Dick need work on recognizing final consonant sounds in words. It is feasible for the teacher to work with these children in a small group until they have mastered this skill.

agree	___:___:___:___:___:___	disagree
ineffective	___:___:___:___:___:___	effective
challenging	___:___:___:___:___:___	unchallenging
disorganized	___:___:___:___:___:___	organized
practical	___:___:___:___:___:___	impractical
fair	___:___:___:___:___:___	unfair
inefficient	___:___:___:___:___:___	efficient

Mark each scale in terms of the effect upon you, the teacher, if this example of instruction were applied in your classroom.

1. Pete and Gary are among the best readers in their third-grade class. It is feasible for the teacher to know that Pete has trouble reading social studies books while Gary who has no trouble with factual material cannot understand non-literal material.

agree	___:___:___:___:___:___	disagree
ineffective	___:___:___:___:___:___	effective
challenging	___:___:___:___:___:___	unchallenging
disorganized	___:___:___:___:___:___	organized
practical	___:___:___:___:___:___	impractical
fair	___:___:___:___:___:___	unfair
inefficient	___:___:___:___:___:___	efficient

2. It is possible for the teacher to know that Dennis is poor in picking out the main idea of a paragraph but good at recognizing all consonant and vowel sounds.

agree	___:___:___:___:___:___	disagree
ineffective	___:___:___:___:___:___	effective
disorganized	___:___:___:___:___:___	organized
practical	___:___:___:___:___:___	impractical
fair	___:___:___:___:___:___	unfair
inefficient	___:___:___:___:___:___	efficient

3. Although Ruth is working in more than one set of materials to learn the short a sound, it is possible for the teacher to know which skill she should be taught next.

agree	___:___:___:___:___:___	disagree
ineffective	___:___:___:___:___:___	effective
challenging	___:___:___:___:___:___	unchallenging
disorganized	___:___:___:___:___:___	organized
practical	___:___:___:___:___:___	impractical
inefficient	___:___:___:___:___:___	efficient

4. It is feasible for a second-grade teacher to use the same 2-1 basal reader with the whole class.

agree	___:___:___:___:___:___	disagree
ineffective	___:___:___:___:___:___	effective
challenging	___:___:___:___:___:___	unchallenging
disorganized	___:___:___:___:___:___	organized
inefficient	___:___:___:___:___:___	efficient

5. It can be expected that a second-grade teacher will know when and how to teach outlining skills to Gary who reads far above grade level.

challenging	___:___:___:___:___:___	unchallenging
disorganized	___:___:___:___:___:___	organized
practical	___:___:___:___:___:___	impractical
fair	___:___:___:___:___:___	unfair
inefficient	___:___:___:___:___:___	efficient

6. It is feasible for Mary Lou, who has not mastered initial consonant sounds, to continue work on them although the rest of the children have mastered this skill and have moved on to new material.

agree	___:___:___:___:___:___	disagree
challenging	___:___:___:___:___:___	unchallenging
practical	___:___:___:___:___:___	impractical
fair	___:___:___:___:___:___	unfair

7. It is feasible in a second-grade classroom to provide Pete with fourth-grade materials which he can read and to give Peggy pre-primer material which is appropriate for her.

agree	___:___:___:___:___:___	disagree
challenging	___:___:___:___:___:___	unchallenging
disorganized	___:___:___:___:___:___	organized
practical	___:___:___:___:___:___	impractical
fair	___:___:___:___:___:___	unfair
inefficient	___:___:___:___:___:___	efficient

8. Marjorie, David, Howard, Dorothy, and several others are working together in a small group on recognizing certain consonant blends. It is possible for the teacher to assess at almost every group meeting which children have mastered this skill and to modify teaching accordingly.

agree	___:___:___:___:___:___	disagree

ineffective   __:__:__:__:__:__ effective
challenging   __:__:__:__:__:__ unchallenging
disorganized   __:__:__:__:__:__ organized
practical   __:__:__:__:__:__ impractical
fair   __:__:__:__:__:__ unfair
inefficient   __:__:__:__:__:__ efficient

9. Jim does not seem to have much interest in reading in the basal reader. The teacher can effectively use non-basal materials to teach him reading skills.

disorganized   __:__:__:__:__:__ organized
fair   __:__:__:__:__:__ unfair
inefficient   __:__:__:__:__:__ efficient

10. Jim, Dennis, Gary, Ruth, and Pete all need to work on the vowel dipthongs oi and oy. It is feasible to meet with this group once or several times, depending on the length of time needed for mastering these sounds in words.

agree   __:__:__:__:__:__ disagree
ineffective   __:__:__:__:__:__ effective
challenging   __:__:__:__:__:__ unchallenging
disorganized   __:__:__:__:__:__ organized
practical   __:__:__:__:__:__ impractical
fair   __:__:__:__:__:__ unfair
inefficient   __:__:__:__:__:__ efficient

11. Gary has mastered all the work taught to the class very quickly. It is feasible to allow him to start working on vowel digraphs even though the rest of the class still is working on consonant blends and short vowel sounds.

agree   __:__:__:__:__:__ disagree
ineffective   __:__:__:__:__:__ effective
disorganized   __:__:__:__:__:__ organized
fair   __:__:__:__:__:__ unfair
inefficient   __:__:__:__:__:__ efficient

MOTIVATIONAL EFFECTS OF COMPETITION

AND GOAL SETTING IN REWARD

AND NON-REWARD CONDITIONS

MARGARET M. CLIFFORD
The University of Wisconsin

ABSTRACT

The effects of six classroom motivational treatments on 112 fifth and sixth grade students were measured using a difference score on a substitution task. Individual goal setting and competitive treatments, under reward and non-reward conditions, are analyzed by means of planned comparisons. Results indicate a significant interaction which suggests caution against an oversimplified interpretation of main effects. A S's performance in a competitive treatment is shown to be dependent upon three factors: his initial ability relative to that of his classmates; the presence or absence of a reward; the homogeneous or heterogeneous nature of the group in competition. The evaluation of this significant 3-way interaction (i. e., ability x reward x grouping) and comparisons of means suggest several tentative hypotheses and raise highly relevant questions regarding the use of classroom motivational techniques which are competitive in nature.

COMPETITION has been examined in the light of such variables as sociometric status (11), rewards and incentives (2, 5, 9), duration, and repetition of a task (1, 14, 15), and inter- and intragroup dependency (3, 10). Studies have been conducted with a variety of Ss, a wide range of tasks, diversity in the use of rewards and incentives, and numerous combinations of treatments. Yet there is a lack of consistency among the findings, and directives for the use of competition in education are very limited. For example, while one study demonstrates that the efficiency of work under the competitive condition is consistently and significantly higher than under the cooperative situation (4), another study reports that performance in a cooperative situation is more efficient than in its contrasting competitive treatment. Phillips and De Vault (12) emphasize that a lack of systematic variation in treatments is a major flaw in cooperation and competition research. Frequently a single competitive treatment is compared with a single noncompetitive treatment and the dichotomy allows for very limited generalization.

The discrepancies in definition of terms also complicate the interpretations and comparisons that might be made among studies. A situation defined as competitive-group versus cooperative-group by one E is identified as group-competition versus individual competition by another (7). For one study, competition is restrictively defined to include only performance in which the success of one member hinders the achievement of other members (3); for another it inclues actively preventing competitors from reaching a goal as well as trying to achieve it for oneself (4). In a much earlier study competition is defined as an effort manifested by one when he is influenced by the " desire to excel "(6).

In examining the effects of competitive techniques in education, it is essential to clearly specify an operational definition and to systematically study treatment variations. The competition in our present educational system takes such various forms as interim reports, contests, class rankings, and scholarship awards. At times recognition is more formal than at others; competitors may be homogeneous or heterogeneous; the competitive nature of the task may be clearly defined or just generally assumed. In any

case, a student's performance is usually compared with that of his classmates, with that of a local or national norm, or, through the use of self-progress charts, it is compared with his previous performance, in which case the student is seen as his own "rival" or competitor.

This study examines the effects of several competitive treatments used for classroom motivation. Competition is defined as a situation in which Ss are encouraged to surpass each other but are unable to directly affect the absolute score of their competitors. The treatments are so designed that three important variables can be studied simultaneously: reward, grouping technique, and ability of Ss. The difference score resulting from a 1 1/2 minute pre- and post-measure on a substitution task is the dependent variable.

DESIGN

A 4x7 randomized block design, consisting of seven treatments and four ability levels, was used. In addition to a control, there were six motivational treatments which varied in type of grouping (i. e., homogeneous grouping, heterogeneous grouping, no grouping) and reward (i. e., absence and presence).

The blocks correspond to four levels of ability as measured by a pretest and are identified as High (H), High Average (HA), Low Average (LA), and Low (L). The dependent variable measured was the difference between the pre- and posttest on a substitution task. Each of the twenty-eight cells in this 4x7 layout contained four observations, yielding sixteen Ss per treatment for a total of 112 observations.

The treatments are identified as follows:

Individual Goal Setting with Reward (I+). Each S's posttest paper had a red line indicating how far he had worked on the pretest and a red-circled item indicating a 10-point increase over his initial score. The circled item was the S's " motivational goal." Each S was promised a reward for achieving this goal.

Individual Goal Setting without Reward (I-). This condition was the same as I+ except that there was no promise of a reward for achievement.

Homogeneous-Group Competition with Reward (Ho+) The S was encouraged to surpass three fellow students all of whom had pretest scores similar to his (i. e., within a 3-point range.) Ss were promised a reward for achieving the highest score in a subgroup.

Homogeneous-Group Competition without Reward (Ho-). This condition was the same as Ho+ except that there was no promise of a reward for achievement.

Heterogeneous-Group Competition with Reward (H e+). The S was encouraged to surpass three fellow students, all of whose pretest scores were quite different and represented a possible range of as much as 25 points. Ss were promised a reward for achieving the highest score in a subgroup.

Heterogeneous-Group Competition without Reward (He-). This condition was the same as He+ except that there was no promise of a reward for achievement.

Control (C). The S worked the posttest as a simple repetition of the pretest.

SAMPLE AND MEASURES

The Ss were 112 students from the fifth and sixth grade team at Empire Elementary School in Freeport, Illinois. The IQ obtained from the school records ranged from 72 to 133; there were seventy-one boys and forty-one girls.

The task used to measure motivational effects of treatments was a digit-letter task resembling the digit-symbol section of the Wechsler Intelligence Test. The task consisted of associating one of six alphabet characters with a 2-digit number as indicated by a key, and reproducing the correct letter in each blank box according to a number printed directly above it. Ss were instructed to work across each row, moving from left to right beginning with the top row and working each in the order in which it appeared. They were informed of the time allowance of 1 1/2 minutes and were permitted to print or write the letter in capital or small form. A task consisted of ninety possible responses.

PROCEDURES

Seven classrooms were used as experimental stations. One E administered all pre- and posttests; Ss remained in their assigned rooms occupied with independent studies while the E moved about testing one group at a time. With each group directions were read, a chalkboard demonstration was given, and questions were answered.

In both the pre- and posttest all students present were allowed to participate, although only 112 Ss were actually used for the final analysis.

For the pretest the E claimed an " interest in knowing how well fifth and sixth grade students performed on a substitution task. " Ss were rank ordered on the basis of this pretest and stratified into four

TABLE 1

PLANNED COMPARISONS

| Comparison | C | I+ | I- | Treatments | | | |
				Ho+	Ho-	He+	He-
1				1	1	-1	-1
2		2	2	-1	-1	-1	-1
3		1	-1	1	-1	1	-1
4	6	-1	-1	-1	-1	-1	-1
5				1	-1	-1	1
6		2	-2	-1	1	-1	1

TABLE 2

MEAN DIFFERENCE SCORES FOR TREATMENTS BY BLOCKS

Blocks				Treatments			
	C	I+	I-	Ho+	Ho-	He+	He-
H	-2.75	1.25	1.75	.75	1.50	-1.25	1.25
HA	-1.75	0.00	.50	7.25	-3.50	-1.25	1.25
LA	.50	1.25	-2.75	4.50	4.00	2.50	.25
L	2.00	2.00	6.00	6.00	-1.75	6.00	1.50
MEANS	-0.05	1.13	1.38	4.63	.06	1.50	1.06

equal-sized blocks corresponding to four ability levels (i.e., High, High Average, Low Average, Low). Twenty-eight Ss were then randomly selected from each of the four ability levels and randomly assigned to one of the seven treatments so that in each treatment there were four Ss from each ability level.

The second task was administered 10 days later under the treatment conditions identified above. For the administration of the posttest each of the seven treatment groups was assigned to one of the seven available stations. In the four competitive conditions subgrouping was necessary in order to create the homogeneous and heterogeneous groups required for the treatments. Thus, the 16 Ss who were assigned to each competitive treatment were further divided into four small groups: for the homogeneous treatments four members from the same ability level formed a subgroup; for the heterogeneous treatments one member from each ability level was randomly assigned to a subgroup. All Ss in competitive treatments were told against whom they were competing and also told the pretest score of each competitor. In the reward treatments Ss were promised candy for successful performance.

ANALYSIS OF DATA

Because the E was interested in particular comparisons among the treatments, the data were analyzed by planned orthogonal comparisons (8, 16).

A total of twenty-four comparisons were used. Six of them were concerned with the differences among treatments; the remaining eighteen were orthogonal polynomial comparisons concerned with block (ability level) by treatment interactions. Table 1 summarizes the six basic planned orthogonal comparisons among treatments.

The first four were designed to answer the major questions with which this experiment was concerned:

1. Is mean performance in the Homogeneous Competition treatments equal to that of the Heterogeneous Competition treatments?

2. Is mean performance in the Competitive treatments equal to that of the Individual Goal

Setting treatments?

3. Is mean performance in the reward treatments equal to that of the non-reward treatments?

4. Is mean performance in the six motivational treatments (both Competitive and Individual Goal Setting) equal to the mean performance of the nonmotivational or control treatment?

The last two comparisons were concerned with reward by treatment interactions:

5. Is the effect of reward the same in the Homogeneous Competition and Heterogeneous Competition treatments?

6. Is the effect of reward the same in the Competitive and Noncompetitive treatments?

RESULTS

The mean difference score for each of the twenty-eight cells is shown in Table 2. Figure 1 presents treatment means across blocks.

The first planned comparison between homogeneous and heterogeneous competition resulted in

FIGURE 1

MEAN DIFFERENCE SCORES BY TREATMENTS
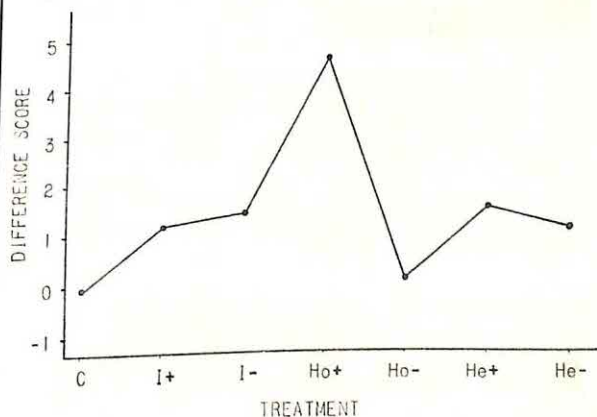
FIGURE 2

COMPETITIVE GROUPING BY REWARD
INTERACTION ACROSS ABILITY LEVELS
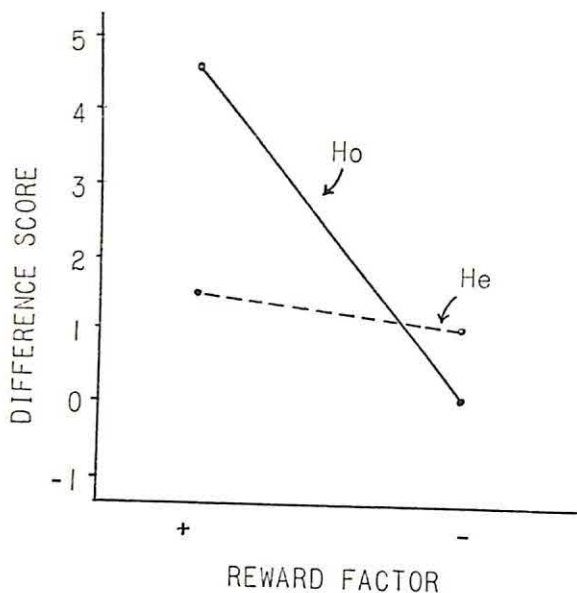
($F(1,84)$ = 1.04; $p \leq .31$). The second planned comparison between Individual Goal Setting and Competitive treatments resulted in ($F(1,84)$ = .39; $p \leq$ .54). Comparison three concerning reward treatments versus non-reward treatments reached the .06 level, ($F(1,84)$ = 2.46). The fourth comparison between control and motivational treatments also resulted in a $p \leq .06$, ($F(1,84)$ = 3.56). Comparison five, used to test interaction among the four competitive treatments across blocks, was found to be significant at the conventional .05 level: (Ho+, He-, versus Ho-, He+)resulted in ($F(1,84)$ = 3.91; $p \leq$ .05). The last comparison, designed to test interaction between the competitive and noncompetitive treatments under the reward and non-reward conditions, resulted in ($F(1,84)$ = 2.32; $p \leq .13$).

Figure 2 represents the interaction between reward and competitive grouping which proved to be significant. Subjects grouped homogeneously responded noticeably better than Ss grouped heterogeneously in the reward condition; the reverse was true in the non-reward condition.

Only one of the eighteen orthogonal polynomial tests concerning the interaction of the six basic treatment contrasts and blocks (i. e. , ability levels)was significant; this test resulted in ($F(1,84)$ = 6.21; $p \leq .01$). It indicated that the difference between the four ability levels in the effect of reward in homogeneous and heterogeneous competition is significant and follows a cubic trend.

Figure 3 shows the reward by grouping interaction for each of the four ability levels. The results represented in Figures 2 and 3 indicate that in predicting treatment performance of fifth and sixth

grade students in a competitive situation, three variables should be considered: presence or absence of a reward, S's ability in relation to classmates' ability, and the homogeneous or heterogeneous nature of competitors.

DISCUSSION

The results of this study clearly indicate that dichotomies such as competition versus noncompetition, reward versus non-reward and goal-setting versus no goal-setting reflect naive simplifications of the group motivation problem in an educational setting. Not one of the first four planned comparisons used to test such global distinctions was significant. The use of systematic variation in competitive treatments, as suggested by Phillips and De Vault (12), has definitely emphasized the complexity of the classroom motivation problem.

It seems relatively safe to conclude that, in general, Homogeneous Competition with reward is the most effective of these seven treatments when used in a classroom situation. This is consistent with much of the competition research as well as literature on the use of rewards and incentives. It is likewise reasonable to conclude, on the basis of this study, that the use of reward in competitive conditions has very different effects dependent upon the homogeneity or heterogeneity of competitors.

If one assumes that the procedures in this study were successful in making Ss aware of the similarity or dissimilarity of their competitors, and if one also assumes that the use of a material reward was perceived as assurance of public recognition for successful performance, one might speculate that the interaction was a result of Ss' discriminating between a socially acceptable and a socially unacceptable victory. Thus, while a S is justified in striving for an award which symbolizes superior performance among equals, it is far less socially acceptable to seek recognition when competitiors are poorly matched on ability—this is particularly true for those who have a marked advantage.

Speculation on each of the interaction patterns for the individual ability levels requires much greater caution; as few as four observations determine each mean in the results diagramed in Figure 3. There is, however, one characteristic of the individual patterns that seems relevant and logically consistent with the "social propriety" speculation: the two higher ability levels show greater resistance to reward in heterogeneous competition than do the two lower ability levels.

Comparing the performance of the four ability levels under homogeneous competition, reveals another striking result: the effect of reward is much greater in the Low and High Average ability levels than in the Low Average and High ability levels. Any interpretation offered on the basis of this limited data is admittedly highly speculative. However, these results raise a question of whether success and/or recognition of success may have relative value dependent upon subgroup status.

In addition to the speculations and tentative hypotheses which can be generated from the results of

FIGURE 3

COMPETITIVE GROUPING BY REWARD INTERACTION BY ABILITY LEVELS
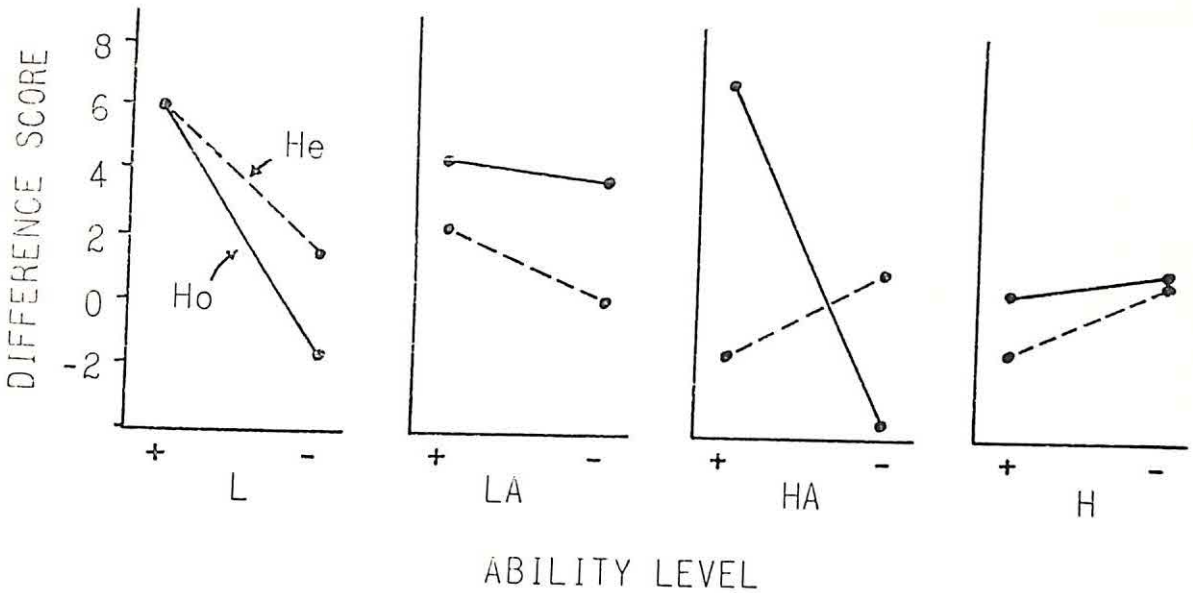
this study, several related questions can be raised. Four which seem to have special relevance for educational motivation are:

1. Is the reward in itself the significant factor or the class recognition implied in obtaining the reward?

2. Do students prefer clearly defined competitive tasks over classroom activities in which competition is assumed but the parameters remain unspecified?

3. To what extent does a S's initial success or failure (relative to his classmates' performance) influence his subsequent performance on a competitive task?

4. Does competition in an educational setting have the same effect on both power and speed tasks?

Replications and carefully designed follow-up studies must be pursued if practical directives for educational competition are to be formulated. Under the present system of education there is little hope of ignoring competition, questionable value in deploring its presence, and no chance of eliminating its influence. On the other hand, it would seem both profitable and practical to research in greater detail the effects of competition both as a prevailing atmosphere resulting from the present educational and cultural patterns and as specific motivational treatments which may be used in classroom situations.

FOOTNOTE

REFERENCES

1. Chapman, J.C.; Feder, R.B., " The Effect of External Incentives on Improvement," Journal of Educational Psychology, 8:469-474, 1917.

2. Davis, R.A.; Ballard, C.R., " The Effectiveness of Various Types of Classroom Incentives," Educational Methods, 12:134-145, 1932.

3. Deutsch, M.J., " An Experimental Study of the Effects of Cooperation and Competition Upon Group Processes," Human Relations, 2:199-232, 1949.

4. Doob, L.W., Social Psychology, Henry Holt and Company, New York, 1952.

5. Gates, Georgina S., " The Effect of an Audience Upon Performance," Journal of Abnormal and Social Psychology, 17: 334-341, 1923-24.

6. Greenberg, Pearl, J., " Competition in Children An Experimental Study," American Journal of Psychology, 44: 221-248, 1932.

7. Hammond, L.K., Goldman, M., " Competition and Non-competition and its Relationship to Individual and Group Productivity," Sociometry, 24: 46-60, 1961.

8. Hays, William, L., Statistics for Psychologists, Holt, Rinehart, and Winston, Chicago, 1963.

9. Hurlock, E.B., " The Use of Group Rivalry as

an Incentive," Journal of Abnormal and Social Psychology, 22:278-290, 1927.

10. Maller, J. B., " Cooperation and Competition: An Experimental Study in Motivation," Teachers College Contributions to Education No. 384 Columbia University Press, New York, 1929.

11. Nelson, J. K.; Johnson, B. L., " Effects of Varied Techniques in Organizing Class Competition Upon Changes in Sociometric Status,"The Research Quarterly, 39:634-639, 1968.

12. Phillips, B. N.; DeVault, M. V., " Evaluation of Research on Cooperation and Competition," Psychological Reports, 3:289-292, 1957.

13. Shaw, M. E., " Some Motivational Factors in Cooperation and Competition," Journal of Personality, 26:155-169, 1958.

14. Sims, V. M., " The Relative Influence of Two Types of Motivation on Improvement," Journal of Educational Psychology, 19:480-484, 1928.

15. Strong, C. H., " Motivation Related to Performance of Physical Fitness Tests, Research Quarterly, 34:497-507, 1963.

16. Walster, G. W., Unequal Cell Frequencies, Arbitrary Weights, and Analysis of Variance, unpublished PhD thesis, University of Minnesota, Minneapolis, 1968.

ACADEMIC ACHIEVEMENT DECLINES

UNDER PASS-FAIL GRADING[1]

RICHARD M. GOLD, ANNE REILLY, ROBERT SILBERMAN, and ROBERT LEHR[2]

State University of New York, College at Cortland

ABSTRACT

College students voluntarily took all their courses or one course on a pass-fail basis. The mean grade point average (GPA) before conversion to pass-fail for freshmen taking all their courses on a pass-fail basis was 1.67 (C-), which is significantly lower than the 2.26 (C+) for controls who wanted but were denied pass-fail grading. Even after returning to conventional grading the former pass-fail students continued to get significantly lower grades than controls. Juniors taking one course on a pass-fail basis received significantly lower grades, before conversion, in their pass-fail course (mean 2.07) than did controls who wanted but were denied pass-fail grading (mean 2.40). There was no compensatory improvement in the grades received in non-pass-fail courses.

ONE PURPOSE of the college experience should be to develop in each student an intrinsic motivation to learn. In actual practice, however, most colleges use extrinsic grades to motivate students to learn. After graduation grades are no longer available, and intellectual activity often ceases.

Low grades and academic dismissal are the punishments for not studying, and high grades and honors are the rewards for successful study. Grade pressure is especially severe for marginal students who want to avoid academic dismissal and for better students who are competing for admission to graduate school. This system encourages students to select courses which promise high grades for a minimum of effort. "The attainment of high grades is often perceived not as a key to success, but as success itself" (2:179). The traditional grading system can also be faulted for its emphasis on information rather than understanding, competition rather than appreciation, and quantity rather than quality. In addition, grades are inconsistent as not all instructors use the same grading standards.

A variety of remedies have been proposed for the presumed undesirable emphasis on grades. At Bennington and Sarah Lawrence Colleges periodic comments by instructors have replaced grades. The effectiveness of this system has unfortunately not been adequately evaluated, but even if it did prove to be a better method of evaluation than the present grading systems, meaningful individualized comments would be impossible with the high student-faculty ratio found at most institutions. An alternative method for de-emphasizing grades that seems more feasible for institutions with high student-faculty ratios is pass-fail grading. In this system, the number of possible grades is limited to two; P for pass and F for fail. Pass-fail grading removes the external motivation for non-failing students and, in theory at least, allows these students to study out of intellectual curiosity, rather than for grades. Pass-fail grading allows a student to de-emphasize, without penalty, aspects of a course or even entire courses that do not interest him. In theory, the student then directs his intellectual efforts to topics that are more consistent with his interests. This may include academic areas in which he would avoid conventionally graded courses for fear of a low grade.

Various forms of pass-fail grading have been initiated at many American colleges. During the fall of 1967 we distributed a questionnaire about grading procedures to fifty-eight selected colleges. Based on newspaper and journal reports, rumor, and pure speculation, it was suspected that these schools had some form of pass-fail grading. In addition, the questionnaire was sent to all fourteen 4-year units of the State University of New York. Of sixty-three replies, thirty-seven (59%) indicated that they did offer some form of pass-fail grading, twelve (19%) others were considering it, and the remaining fourteen (22%) had no provision for pass-fail grading. That pass-fail is a relatively new grading system is

evidenced by the fact that only one of the schools had been offering it for longer than 3 years. A variety of pass-fail grading practices were found to exist. A few schools, including Massachusetts Institute of Technology, Yale, and Antioch, use pass-fail grading exclusively. In all but seven (19%) of the schools offering pass-fail grading, courses that a student was allowed to take on a pass-fail basis were limited to one course per semester and four per college career. Of the schools that offered pass-fail grading, nineteen (51%) offered it to all undergraduates and seventeen (46%) offered it only to upperclassmen. In most schools, pass-fail was available in all courses offered. However, thirty (82%) limited the option to elective courses outside the student's major. At Brandeis University a questionnaire (1) was sent to participating students and faculty. The results indicated generally uncritical support for pass-fail grading.

Our questionnaire asked each school to evaluate the success of its pass-fail grading. Most of the comments on the overall success of pass-fail grading were subjective impressions. Of the thirty-seven schools offering a pass-fail option, seventeen (49%) felt that pass-fail had achieved some degree of success, seventeen (49%) felt that it was too early to tell, and only one (2%) judged the system to be unsatisfactory. The criterion used was student-faculty acceptance, rather than measurable intellectual activity.

The apparent popularity of pass-fail grading seems to indicate that grade pressure is unpleasant. Reducing unpleasant grade pressure may be a desirable goal; however, the academic consequences of pass-fail grading must also be considered. The present study is a controlled evaluation of the effects of both one-course and complete pass-fail grading on academic performance.

METHOD

The Ss were students at Cortland College of the State University of New York. Virtually all were New York State residents.

During the summer of 1967 a stratified sample of 293 entering college freshmen with low (379 to 477), middle (511 to 559), and high (580 to 785) Scholastic Aptitude Test Verbal (SAT-V) scores were matched by SAT-V score and sex and assigned to either group 1, all courses pass-fail; group 2, one course pass-fail; or group 3, a control group. Similarly, a stratified sample of 218 college juniors with low (1.9 to 2.1), middle (2.2 to 2.5), and high (2.8 to 3.9) GPA's as of June 1967, were matched for GPA and sex and then assigned to group 2, one course pass-fail; or group 3, a control group. No juniors were assigned to group 1. All students were notified during the summer that they had been selected to participate in a 1-year pass-fail study and were requested to attend a meeting on the first day of the fall semester. At this meeting they were told the purpose of the study, and unaware of which group he had been assigned to, each freshman indicated (1) whether or not he would accept an all course pass-fail option if given the opportunity; and (2) whether or not he would accept a one course pass-fail option, and if so, in which course he would use the option. There were no restrictions on which course could be

selected. Juniors were offered only the one course pass-fail option. Students were told that their choices would be binding. Finally, students were informed as to which experimental group they had been assigned. Instructors were not told which of their students, if any, were to receive pass-fail grades. During the semester, the students taking pass-fail courses received feedback, in the form of examination grades, as did their classmates. After final grades were submitted by instructors at the end of the semester, the appropriate A through D grades were converted to P (Pass), and D- and E grades were converted to F (Fail). These are the only grades appearing on students' transcripts. P or F grades were not used in computing grade point averages, but P grades were credited toward graduation. The traditional A through E grades of pass-fail students were used for evaluation purposes only. The traditional grades submitted for students under each pass-fail condition were compared with grades for the control group students that had requested the same pass-fail option but were not allowed to have it.

It was hoped that under pass-fail grading learning experiences might tend to be oriented away from rigid compliance with course assignments or cramming for examinations. This nonconformity could produce an initial deterioration in grades. However, insofar as a college education is cumulative, these independent learning experiences, if they occur at all, should eventually lead to improved academic performance as the accumulated wisdom is related to new courses. To test for such a delayed effect, the grades of the all-course pass-fail group were studied for the first semester in which they returned to traditional grading (1968) and again for the first semester of the Junior year (Fall 1969).

RESULTS

Despite the supposed evils of conventional grading, only 28 percent of the freshmen wanted to take all their courses on a pass-fail basis, while 78 percent of the freshmen and 80 percent of the juniors wanted to take one pass-fail course.

Instructors submitted A through E ($\pm$) grades for all students and, where appropriate, the registrar made conversions to P or F grades.

All Courses Pass-Fail

The data for the all-courses pass-fail groups are presented in Table 1 and Figure 1. The mean GPA before conversion for freshmen taking all their courses on a pass-fail basis was 1.67 (C-), as compared with 2.26 (C+) for the freshmen controls who wanted but were denied the same option ($p < .01$, t-test). The difference between the all-course and control grades was greatest for the subgroup with high SAT-V scores. To summarize, all course pass-fail grading led to a decline in academic performance.

Academic performance during the first semester after the all-course pass-fail experience is also presented in Table 1. These follow-up averages do not include data from seven experimental and three control students who, after poor academic performance, withdrew or were dismissed from the college without completing 1 semester under conventional grading. In this follow-up comparison both groups

TABLE 1

MEAN GRADE SUBMITTED FOR COLLEGE
FRESHMEN TAKING ALL COURSES ON A PASS-
FAIL BASIS

FIRST PASS-FAIL SEMESTER GRADES

	Experimental		Control	
SAT Verbal Score	GPA	N	GPA	N
580-785	1.55	9	2.53*	8
511-559	1.36	7	2.14	13
379-479	1.91	13	2.15	6
All Ss	1.67	29	2.26*	27

FIRST FOLLOW-UP SEMESTER WITH
CONVENTIONAL GRADES

All Ss	2.28	22	2.72*	24

SECOND FOLLOW-UP; FALL SEMESTER
OF JUNIOR YEAR

All Ss	2.68	18	2.85	20

* $p < .01$

received conventional grades. The only difference
between the groups was the previous pass-fail expe-
rience of the experimental group. Mean GPA was
2.28 (C+) for the pass-fail group and 2.72 (B-) for
the controls ($p < .01$, t-test). Thus, taking all
courses on a pass-fail basis for 1 or 2 semesters
also impaired subsequent academic performance un-
der traditional grading. One year later, in the first
semester of their junior year (Fall 1969), the mean
GPA was 2.69 for the pass-fail group and 2.86 for the

FIGURE 1

MEAN OF GRADES SUBMITTED FOR FRESHMEN
TAKING ALL COURSES ON A PASS-FAIL BASIS.
CONTROLS WANTED BUT WERE DENIED PASS-
FAIL GRADING
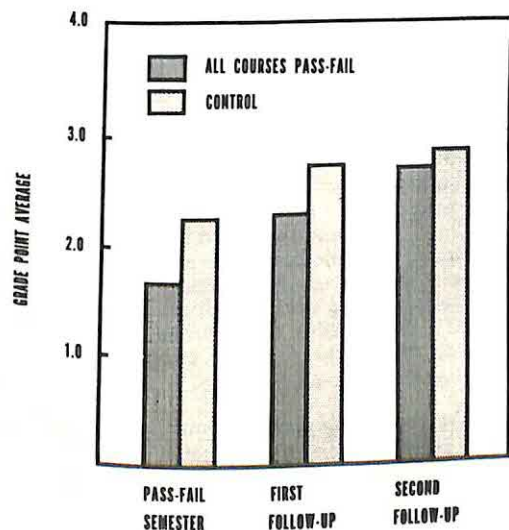
FIGURE 2

MEAN OF GRADES SUBMITTED FOR FRESHMEN
AND JUNIORS TAKING ONE COURSE ON A PASS-
FAIL BASIS. CONTROLS WANTED BUT WERE
DENIED PASS-FAIL GRADING
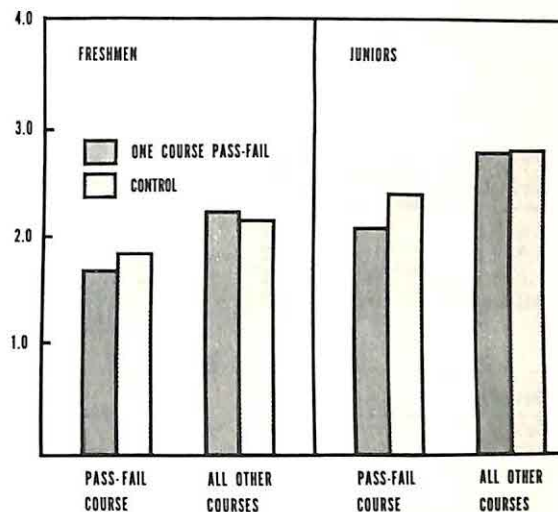
controls ($p > .05$, t-test). These averages do not
include data from eleven experimental and seven con-
trol students who withdrew or were dismissed from
the college during the study.

One Course Pass-Fail

Mean grades for the one-course pass-fail option
are presented in Table 2 and Figure 2. For all ranks,
the mean A, B, C, D, E ($\pm$) grade submitted for the
pass-fail course was lower than the mean grade for
controls in the course they wanted but were not per-
mitted to take on a pass-fail basis. The difference
was significant for juniors alone and for freshmen
and juniors combined ($p < .05$, t-test), but did not
reach significance for the freshmen alone. Grades
in the non-pass-fail courses, that is, all courses ex-
cept the pass-fail choice, showed no significant dif-
ferences between experimental and control Ss. Thus,
the one course experimental group students failed to
show significant compensatory improvement in their
non-pass-fail courses.

Within Group Comparisons

With data for freshmen and juniors combined, the
students taking one pass-fail course received signif-
icantly lower grades in their pass-fail course than
their non-pass-fail courses ($p < .001$, sign test).
However, the control students also received signif-
icantly lower grades in the course they wanted but
were denied the opportunity to take on a pass-fail
basis ($p < .001$, sign test).

DISCUSSION

Student Acceptance

The high level of student acceptance of limited
pass-fail grading reported by other institutions was
reflected in the present study, with 78-80 percent of

TABLE 2

MEAN GRADE SUBMITTED FOR STUDENTS TAKING ONE COURSE ON A PASS-FAIL BASIS

| | FRESHMEN | | | | | | | |
| | Pass-Fail Course | | | | | Non Pass-Fail Courses | | |
SAT Verbal Score	Experimental	N	Control	N	p	Experimental	Control	p
580-785	1.85	25	2.22	20	n.s.	2.53	2.43	n.s.
511-559	1.69	24	1.85	24	n.s.	2.33	2.15	n.s.
379-479	1.47	22	1.48	21	n.s.	1.83	1.91	n.s.
All Ss	1.67	71	1.83	65	n.s.	2.23	2.15	n.s.
Cumulative GPA				JUNIORS				
2.8-3.9	2.54	26	2.67	16	n.s.	3.25	3.30	n.s.
2.2-2.5	1.92	31	2.43	37	<.05	2.72	2.76	n.s.
1.9-2.1	1.83	32	2.23	33	<.05	2.46	2.60	n.s.
All Ss	2.07	89	2.40	86	<.05	2.78	2.80	n.s.
	FRESHMEN AND JUNIORS COMBINED							
All Ss	1.89	160	2.15	151	<.05	2.54	2.52	n.s.

the students electing to take one course on a pass-fail basis. However, pass-fail grading was intentionally made attractive in order to entice a large number of students to participate in the study. If a student passed a pass-fail course with a low grade, he received full academic credit, but the low grade was not averaged into his GPA. If he failed the course he naturally received no academic credit, but the failure still was not averaged into his GPA. In order to benefit from the option, the student had only to be able to select a course in which he would get a low grade. Grades achieved by the control groups in the one course they wanted, but were denied the opportunity to take on a pass-fail basis, were in fact lower than the grades in their other courses. Thus, students elect as pass-fail choices courses in which they anticipate low grades. Thus, if students receive low grades in pass-fail courses in uncontrolled studies, this may merely be due to their skill in selecting courses in which they would perform poorly anyway.

Pass-fail grading reduces grade pressure and thereby permits the student to divert some of his energy away from grade-oriented studying. Is this released time used to pursue additional academic interests, or does the overall level of academic activity decline to the minimum necessary to obtain the pass grade? The follow-up grades for the all-course group provide a partial answer to this question. If during the all-course pass-fail experience students had spent test preparation time on intellectual pursuits, then that activity should have been reflected in the following year's grades. There was no evidence that a year of all-course pass-fail grading was subsequently advantageous. On the contrary, during the first follow-up semester on conventional grades the all-course group earned significantly lower grades than the control group that had wanted but had been denied the same pass-fail option. For the second follow-up year the results were in the same direction but were not statistically significant.

Somewhat similar results were obtained with

one-course pass-fail grading. Compared to the control groups, juniors taking one pass-fail course obtained lower grades in that course, while their performance level remained unchanged in their other courses. Thus, the data suggest that time diverted from the one pass-fail course was not spent on regularly graded courses.

In order to have the appropriate control groups, the students participating in this study could not be told which experimental group they were in until after they completed academic registration. However, one of the theoretical advantages of pass-fail grading is that it encourages students to take courses they would not otherwise be willing to attempt. In order to facilitate and at the same time evaluate the use of this advantage, we had a pass-fail table on registration day for students who would take a particular course if they had a pass-fail option but would take a different course if they were in the control group. These students were to obtain class cards to cover both possibilities, after which we would immediately tell them which group they were in so that they could complete their registration. Out of the 511 students involved in the study, only one indicated that he would take a different course if he had a pass-fail option. Thus, pass-fail grading did not encourage students to take more challenging courses.

The academic decline observed under pass-fail grading in this study may be attributed to the students' previous experience. To students who have been extrinsically motivated throughout their high school education, pass-fail grading may represent only an escape from serious study. For this reason, pass-fail grading might prove more beneficial if instituted earlier in the student's career, before grade motivation becomes an obstacle.

Pass-fail grading probably will not solve all problems of academic evaluation, but if initiated in such a way as to minimize its abuses, it is a step toward alleviating grade pressure. It may even offer students the opportunity to develop an intrinsic

motivation to learn. The present study did not adequately measure this change, if it occured, since extrinsic grades were the s o l e evaluative criterion. However, there was no evidence in the follow-up data of any long term benefits. If pass-fail is to be implemented at all, limits should be placed on the option. The advantage of reduced grade pressure may appear to some to outweigh the decline in performance in courses such as terminal electives not in the student's major.

In conclusion, the data suggest that what appears to be a trend toward pass-fail grading may be unwarranted. Students have learned how to w o r k f o r grades and appear to learn a little in the process. It is as yet doubtful whether many have discovered how to learn without grades.

FOOTNOTES

REFERENCES

1. Sgan, Mattew R., "The First Year of Pass-fail at Brandeis University: A Report," Journal of Higher Education, 40:135-144, 1969.

2. Stallings, W. M.; Smock, H. R.; Leslie, E. K., "The Pass-fail Grading Option," School and Society, 96:179-180, 1968.
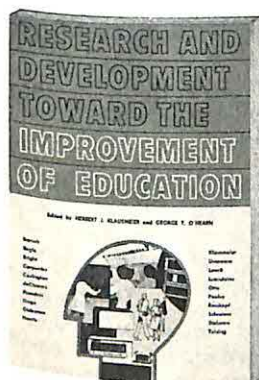
EGRULE VS RULEG TEACHING

METHODS: GRADE, INTELLIGENCE,

AND CATEGORY OF LEARNING

G. D. HERMANN
Macquarie University, Sydney, Australia

ABSTRACT

Methodological problems have limited the usefulness of findings from experiments into learning by discovery. By using programmed instruction materials, a within-class design, and other controls, an attempt was made to remove confounding. Two tasks were used: concept learning and principle learning. For each task, a separate 2x2x2 factorial design containing sixteen Ss in each cell was used. Independent variables were instructional method (egrule and ruleg), school grade (9 and 5), and intelligence (high and average). A set of eight different measures, involving retention, transfer, and ease of relearning, was used for each task. It was found that the egrule and ruleg methods did not differ significantly, and that interaction between instructional method and the other variables was low.

IN EDUCATIONAL literature relating to the teaching of mathematics and science, the present fashion is the advocacy of learning by discovery as the most effective teaching method. This belief is not based on firm experimental evidence. In 1966 Wittrock wrote:

> Many strong claims for learning by discovery are made in educational psychology. But almost none of these claims has been empirically substantiated or even clearly tested in an experiment (26:33).

Later reviews support this statement (10, 22). Methodological problems relating to a lack of careful specification of the treatments used and to a lack of control of possible confounding variables prevent the unambiguous interpretation of the results from most experiments.

In a large number of studies, nonsignificant differences have been obtained. The recent experiment conducted by Tanner (23) is an example. Ninth—grade pupils (N = 360) were taught the principles of mechanics in three different groups: expository-deductive, discovery-inductive, and unsequenced-dis-

covery. Using a variety of measures, including both transfer and retention, and a within-class design involving programmed instruction in an attempt to control confounding variables, it was found that there were no significant differences among the three methods. In this experiment, however, as in a number of other experiments into learning by discovery, the actual degree of learning was low, making interpretation of the results more difficult.

It is possible that the insignificant findings represent the true position, and that students learn equally well from ruleg and discovery methods; significant results could be due to confounding. Alternatively, one method may be generally superior, but more significant results have not been obtained due to such factors as inappropriate measure, lack of definition and specificity of treatment, and lack of control; or possibly one method may be appropriate to certain subsets of learners, which have not yet been identified because no appropriate experiments have been undertaken. In the present study, an attempt was made to achieve a much higher degree of control. This was done through the use of programmed instruction booklets in which the two instructional methods had exactly the same verbal

content, organization, manual activity, and so on. Approximately equal numbers from each school class were in the two instructional groups, to control for previous types of learning experiences. Subjects were screened to ensure that they had no prior knowledge of the task. Subjects in both groups had the same time allocation, and were presented with items at a fixed rate in order to ensure that all Ss attempted all items. Efforts were made to optimize the learning for each group within this time allocation. Other sources of confounding were also eliminated.

In very few experiments have operational definitions of the discovery and ruleg learning methods been used. In the present experiment, the egrule (examples followed by rule) and ruleg (rule followed by examples) methods were clearly specified; the only difference was the placement of the rule.

It has been argued (10) that it is necessary to study the interaction of variables, since it is possible that learning by discovery is more relevant to certain population subsets than to the universal set. In addition to instructional method, the variables investigated were school grade, intelligence, and category of learning. School grade was selected since it was related to one of the few predictions concerning learning-by-discovery variables. Ausubel postulated that ruleg methods were generally superior to discovery methods on a time-cost basis; he recognized, however, the greater relevance of learning by discovery for:

> ... children approximately below the age of twelve; ... during the elementary-school years; ... for children who are still functioning at Piaget's level of concrete operations (1:23).

It was hypothesized, therefore, that on a new task, elementary school children would perform relatively better on the egrule method, and that high school children would perform relatively better on the ruleg method.

Although some studies have been reported in which lower ability groups tended to learn relatively better from discovery methods than did high ability groups (10), this finding has not always been obtained (14, 23). It was hoped that the grade x intelligence x method interaction in the present experiment would clarify this inconsistency.

Concerning category of learning, Gagné has stated: "The learning of concepts appears to require a process of discovery..." while "Principle learning can be done with or without discovery" (7: 149-150). Since, by definition, it is not possible to use precisely the same subject matter for both concept and principle learning, it was necessary to use two different tasks; any difference involving the two tasks could thus be ascribed to factors other than category of learning (e. g., difficulty).

It seems necessary to justify the specific selection of egrule and ruleg methods. There are a number of methodological problems related to the presentation of a rule following learning. Retroactive inhibition, the Zeigarnik effect (of superior mem-

ory for unfinished tasks), and the Ovsiankina effect (of resumption of incomplete tasks) may possibly be of importance depending on the operations involved in the learning sequence, and must be taken into account.

There seem to be five approaches with respect to this cluster of issues:

(1)   To present examples only (i. e., no rule) to the discovery Ss. This eliminates the possibility of retroactive inhibition affecting only the discovery group; two confounding problems are, however, introduced. First, is the confounding problem with respect to possible additional practice time for one group only, as reported by Kersh (12, 13). Second, and possibly related to the first, is the opportunity for the Zeigarnik and Ovsiankina effects to operate.

(2)   To use a much longer learning session. This would probably have the effect of reducing "motivated practice" between the learning and testing sessions; Cronbach (4) quotes an experiment conducted by Kersh involving sixteen training sessions in which it was found that there was no difference between groups in the use of information outside the class. The problem of less control for long-term experiments with human Ss still remains.

(3)   To present the rule at the end of the learning session to the inductive group only. This introduces confounding, since only the egrule group would suffer the possibility of retroactive inhibition effects.

(4)   To present the rule at the end of the learning session to both groups. There would seem to be the possibility that the retroactive inhibition effect could act differentially with respect to the different groups; for example, during the learning session the ruleg group could have been operating in terms of a "rule," while the egrule group could have been operating in terms of a "method."

(5)   To present the rule to the discovery group about half-way through the learning session. This would, it is hypothesized, considerably reduce the effects of retroactive inhibition, since there is time for the rule to be assimilated while the Ss work through the remainder of the examples. Other effects would appear to be eliminated since the rule is presented to both groups.

The last approach was the preferred one, and so the egrule group was first presented with the rule (in each unit) just after they were half-way through the unit. This was achieved by developing the units as two modules. For example, (where R = Rule; E = Example; and A = Answer)

Module 1:   R + E + A;   E + A;   R + E + A;   E + A.

Module 2:   E + A;   E + A;   E + A;   E + A.

For each unit, the sequences were:

Ruleg group: Module 1-Module 2.
Egrule group: Module 2- Module 1.

The ruleg and egrule sequences were chosen since they represent the two most common methods

TABLE 1

MEAN AGES (YEARS AND MONTHS) AND IQ'S WITH RESPECT TO GRADE, INTELLIGENCE, INSTRUCTION-AL METHOD, AND CATEGORY OF LEARNING

| TASK | | | Grade 9 | | | | Grade 5 | | | |
| | | | High I Q | | Average I Q | | High I Q | | Average I Q | |
			Egrule	Ruleg	Egrule	Ruleg	Egrule	Ruleg	Egrule	Ruleg
CONCEPT (Matricu-lation)	Age	Mean	14-11	15-3	15-3	15-5	10-7	10-9	10-11	11-0
		S.D.	0-6	0-3	0-4	0-6	0-4	0-4	0-5	0-5
	I Q	Mean	123	119	102	101	119	120	104	103
		S.D.	6.5	5.6	7.1	7.4	6.5	5.7	4.3	5.9
PRINCIPLE (Map-Read-ing)	Age	Mean	15-0	15-0	15-3	15-3	10-9	10-8	11-0	11-1
		S.D.	0-3	0-5	0-5	0-5	0-3	0-4	0-4	0-5
	I Q	Mean	124	121	98	101	118	120	102	101
		S.D.	5.1	6.1	6.7	5.5	3.8	5.3	5.1	6.3

of teaching, viz the ruleg and the inductive methods. The specter of confounding was removed since, in addition to the resolution of the methodological problems raised above, there was only one difference between the groups, namely the stage at which the rule was given. It should be noted that in Ausubel's terminology, both methods were reception learning, since the rule was given in each case.

The objective of this experiment was to conduct a controlled investigation into the interaction of variables in egrule and ruleg methods. For each of two separate tasks (concept and principle learning), a 2x 2x2 factorial design was used; the independent variables were instructional method, school grade, and intelligence. Various retention, transfer, and relearning posttests constituted the dependent variables and were taken 4 weeks after the learning session.

METHOD

Subjects

For each task, 16 Ss were tested in each cell (a total of 256 Ss). There were approximately equal numbers of males and females. Grades 5 and 9 were selected for the following reasons:

(1) they are in the elementary and high schools respectively,
(2) they correspond to two Piagetian stages, and
(3) they appear to cover the maximum grade span for which suitable learning tasks could be devised.

Since it was possible that geography students could have studied relevant material, non-geography students only were used as grade 9 Ss. Each S attempted one task only.

The I Q ranges were 85-110 and 111+ for the average and high intelligence groups respectively. School records were used to obtain the I Q data. Details concerning the groups are shown in Table 1.

Materials

The learning and testing materials were devised by the author, and were extensively pretested before use. They consisted of linear programmed instruction booklets and separate answer sheets, while a tape recorder was used to ensure a fixed rate of progress through the booklets. Hints and fading techniques were used (for both methods) where relevant. Both tasks were meaningful and non-arbitrary.

Principle learning task. A map-reading task was used. The four principles (listed in Table 2) can be rearranged in the form of Gagné's (6:58) "if...then" paradigm.

Each item in the booklet was written on the right hand page; a dash indicated that a word or number was to be constructed and written on the answer sheet. The verbal content of the items was kept to a minimum. Opposite the item was the appropriate map, while the correct answer was on the back of each question page. To illustrate the task, map 2 and its items (24-31) are presented in the Appendix. A scale, consisting of printed black lines on a small piece of Perspex © , was used by Ss to interpolate between the grid lines.

Concept learning task. The concept used was " matriculation to Sydney University," which can be considered as a classification rule. The four defining characteristics of matriculation (see Table 2) can be shown as a decision tree (eg., 11); also, this concept would appear to be consistent with Gagné's

TABLE 2

SUMMARY OF ORGANIZATION OF THE
PROGRAMMED INSTRUCTION BOOKLETS

Concept   Principle

INSTRUCTIONS

Items 1-3

These items are to give Ss experience in the lay-
out of the programmed instruction booklets and the
answer sheets.

Some basic
terminology is
introduced in
these items.

PRELIMINARY SCREENING TEST

Two items similar to those at the end of the pro-
gram. No answers are given.

REVISION OF INSTRUCTIONS

LEARNING

Items 4-47

4-11: Number of exam subjects   4-23: Easting
12-23: Levels of exam subjects   24-31: Northing
24-31: English compulsory   32-39: Six-digit
                                            position
32-47: Agriculture or Mathe-   40-47: Principle
       matics or Science com-   for third
       pulsory   digit be-
                                            ing zero

CHECK ON LEARNING

48-49: Two questions with no answers given, as
       a check on whether S had simply been copy-
       ing down the correct answers

definition (6:58). The task could be described as
one of concept attainment (see 2).

A tabular layout was used. At the top of each
(left hand) page was the background material neces-
sary for Ss to appreciate the type of criterion being
taught. (The right hand pages were blank.) The cri-
teria were related to the column headings, while each
row represented one item (i. e., a list of examination
subjects passed). Subjects recorded each answer
(always a combination of "yes" and/or "no") in
an answer booklet resembling the learning booklet;
the correct answer was given on the next (left hand)
page.

A summary of the organization of the programmed
instruction booklets for each task is presented
in Table 2. For each individual task, the examples,
verbal material, etc., are exactly the same; the only
difference is the placement of the rule. The
results on the learning check items are
presented in Table 3.

Posttests. Posttests A and C (different sets of
tests for each task) were parallel forms. Each test
contained about 50 percent retention, 25 percent
near transfer, and 25 percent far transfer items.
The retention items were exactly the same as some
of the items in the original learning task and, for
posttest C, were exactly the same as some of the
items in the relearning program. The far transfer
items generally required Ss to perform operations
which were the reverse to those required in the
learning process; for example, whereas in the learn-
ing of the map-reading task, Ss were required to
read the position from a map, in the far transfer
task they were given a position and asked to place it
on a map. All answers had to be constructed; the
scoring was based on the correctness of the answers.

Relearning programs constituted posttest B.
Omitting the three introductory items, alternate
items of the remaining forty-four items of the rel-
evant original learning program (i. e., 22 items)
were used. The arrangement of the learning pro-
grams meant that items containing the rules were in-
cluded (i. e., those items with even numbers). Both
egrule and ruleg relearning programs were pre-
pared. The answer to every second item was omit-
ted, thus resulting in eleven test items. For the re-
learning programs, two measures were employed;
first, the correctness of the constructed answers
for the eleven items without answers given, and
second, the time taken to complete the program.
Separate answer sheets were used. There was no
time limit on the tests individually or as a whole;
the time, which was written on the chalkboard every
minute, was copied down by Ss onto their answer
sheets at the commencement and at the end of each
of the three tests.

That the various tests were effective was dem-
onstrated by the data showing that on each of the
fourteen tests used, at least two significant results
were obtained at the .01 level (with one minor excep-
tion). It was therefore decided to undertake item
analyses on only one set of parallel forms: Posttest
A. Each analysis was performed on eighty scripts,
ten scripts being randomly selected from each of the
eight categories. Mean difficulty ranged from 0.46-
0.55 and 0.51-0.71 for the matriculation and map-
reading tasks respectively (where 1.0 = all items
answered correctly). Kuder-Richardson 20, (KR-20)
reliabilities ranged from 0.77-0.96. Discrimination

TABLE 3

NUMBERS OF SUBJECTS PASSING THE LEARNING
CHECK ITEMS

Task	Instructional Method	Grade 9		Grade 5	
		High IQ	Average IQ	High IQ	Average IQ
Matric- ulation	Egrule	15	9	4	2
	Ruleg	16	11	6	5
Map- Reading	Egrule	16	15	12	10
	Ruleg	16	15	13	10

indices (phi and point-biserial) were satisfactory.

Procedure

The experiment was undertaken in schools, the class group being the unit for learning and testing. At the elementary school level, nine classes at four schools were involved. At the high school level, ten classes of non-geography students at six schools were involved. It was anticipated that the possibility of communication would be decreased by having a low number of classes at any one school. For all elementary school classes, the learning and testing sessions were held in the morning.

For control purposes, it was considered that all Ss in the same grade should attempt all items in the same time allocation. A fixed rate of presentation was thus used; available data from programmed instruction experiments tend to indicate that there is no difference in attainment if a fixed rate of presentation, as compared to S's own rate, is used (15, 20). Following pretesting, when it was found that the appropriate time for grade 5 Ss was too long for grade 9 Ss, who became frustrated or bored, different time allocations were utilized (43 and 30 minutes respectively, not including instructions).

To control for background knowledge the (few) Ss who passed the preliminary screening test were excluded from the analysis. The final sample thus consisted of Ss who were randomly assigned on a within class basis. They attended the learning and testing sessions, had IQ data available, and failed the screening test.

Learning session. The pupils were seated in their normal pattern. E was introduced by the principal or class teacher, who stated that E was interested in the development of new teaching methods; no mention was made of discovery learning or of subsequent testing.

Before the material was handed out, E asked the Ss to leave it unopened on their desks. The learning programs were distributed in such a way that approximately the same number of pupils were placed in the various cells formed from the instructional method and category of learning variables. In order to remove any tendency to "cheat," the matriculation and map-reading programs were handed out to alternate columns of Ss. To overcome any bias due to class arrangements, Ss in each column were alternately given egrule and ruleg programs.

Ss were asked to write their names on the answer sheets. E then stated that there were two separate tasks, and that each person would do one task only. The matriculation task was identified, and the definition of matriculation was read from the front of the booklet. Using sheets of cardboard, tied together and otherwise set up to resemble the first pages of the programmed instruction booklet, E identified the following: position of question (and the red box surrounding it), position of question number, columns and rows constituting the question, position of answer on subsequent page (together with its green box), and position of next question. E told Ss that the answer to each part of each question was either "yes" or "no." Again, using sheets of cardboard, E identi-

fied the following for Ss doing the map-reading task: position of map, position of question, position of question number, and position of answer on next page. E emphasized to Ss that all answers were to be written on the answer sheets. They were told to read the printed instructions to themselves while E read them aloud. The instructions were:

An attempt is being made to improve teaching methods, and your help will be greatly appreciated.

In this booklet are a number of questions. Work out the answer, write it down in the space provided, then see if it is exactly the same as the correct answer given on the next page.

If your answer is different, or if you haven't worked out the answer before the bell, quickly try to work out how to get the correct answer.

Go on to the next question when the number is called, even if you have not worked out the answer; the following questions will help you.

If you finish before the number is called, think about the method for getting the correct answer; do not go on to the next question until the number is called.

Do the best you can.

A further instruction was added. "Commence when you hear the number 1 on the tape recorder." E then pressed the appropriate button on the tape recorder. On the cassette were the numbers at the intervals specified. Immediately before each number was a warning bell. Depending on the time allocation for the item, another bell sounded 7 1/2 to 15 seconds before the warning bell for the next item (to encourage students " to try to work out how to get the correct answer " as stated in the instructions).

Immediately before the sounding of " four, " the tape recorder was held, and after asking Ss to read the following printed instructions to themselves, E read them aloud, as follows:

We will stop to make sure that everyone knows exactly what to do.

1. (a) When you have worked out the answer, write it down.
    (b) Look at the correct answer, and see if your answer is exactly the same.
    (c) If your answer is different, quickly try to work out the method for getting the correct answer; do not change your original answer.
    (d) Go on to the next question when the number is called.

2. (a) If you haven't worked out the answer when the bell sounds, look at the correct answer and quickly try to work out the method for getting the correct answer.
    (b) Go on to the next question when the number is called.

3. Don't worry if your answer is wrong; there

TABLE 4

EXPERIMENTAL RESULTS

Measure		Grade 9				Grade 5			
		High IQ		Average IQ		High IQ		Average IQ	
		Egrule	Ruleg	Egrule	Ruleg	Egrule	Ruleg	Egrule	Ruleg
Matriculation (C)									
A: Retention (32)*	Mean	28.69	26.25	17.63	20.06	16.69	21.88	17.75	14.94
	S.D.	3.96	4.49	7.83	8.44	6.77	4.82	5.91	7.66
A: Far Transfer (19)*	Mean	15.25	15.25	11.19	13.06	8.69	10.38	8.88	8.81
	S.D.	3.69	3.92	3.94	3.97	4.54	4.54	3.50	5.10
A: Near Transfer (20)*	Mean	16.50	16.38	12.63	14.13	9.63	9.88	7.69	8.00
	S.D.	1.10	1.15	4.40	3.10	4.47	5.14	3.34	3.33
B: Correct Responses (30)*	Mean	29.38	28.31	23.56	25.38	23.94	24.63	19.00	19.56
	S.D.	1.09	3.67	5.24	5.75	4.27	4.95	4.50	5.34
B: Time	Mean	8.06	9.25	11.31	10.50	18.75	17.13	17.06	13.31
	S.D.	2.27	2.05	4.09	3.18	8.67	3.42	6.71	4.01
C: Retention (32)*	Mean	31.31	30.06	26.25	28.63	26.94	25.94	19.00	22.00
	S.D.	1.74	2.98	5.01	3.88	5.17	4.67	6.08	5.94
C: Far Transfer (19)*	Mean	17.13	16.38	11.63	12.38	10.31	12.88	7.19	7.75
	S.D.	2.55	3.74	3.98	4.77	5.64	4.41	3.71	5.75
C: Near Transfer (20)*	Mean	14.88	15.63	12.00	13.75	11.19	11.75	8.00	8.06
	S.D.	2.92	1.78	3.72	3.24	3.25	3.32	3.45	3.04
Map-Reading (P)									
A: Retention (47)*	Mean	41.56	41.13	30.06	34.75	24.25	30.06	18.88	21.81
	S.D.	7.59	7.46	13.90	10.43	14.45	15.37	15.72	14.95
A: Far Transfer (26)*	Mean	21.38	20.06	12.25	11.00	7.88	7.69	5.44	3.13
	S.D.	7.04	6.79	9.88	8.12	8.26	6.50	5.11	5.20
A: Near Transfer (23)*	Mean	21.88	20.13	15.94	18.63	13.25	16.44	10.00	10.81
	S.D.	1.78	4.62	8.58	5.70	6.67	7.56	8.45	7.35
B: Correct Responses (53)*	Mean	51.00	50.81	45.50	46.06	45.94	48.06	32.94	33.44
	S.D.	1.90	2.01	12.40	8.13	8.05	5.21	15.18	16.08
B: Time	Mean	9.38	8.13	10.81	11.13	16.44	16.06	17.63	16.25
	S.D.	1.89	2.42	3.10	2.63	2.94	4.74	7.08	4.87
C: Retention (46)*	Mean	44.56	44.13	38.63	38.25	33.94	41.38	25.31	29.00
	S.D.	1.32	2.73	12.58	9.16	13.16	8.40	14.66	15.38
C: Far Transfer (26)*	Mean	24.13	22.44	15.50	12.69	11.00	12.75	5.00	2.94
	S.D.	3.74	6.13	10.15	9.21	7.93	8.40	4.95	5.23
C: Near Transfer (23)*	Mean	21.88	20.94	19.06	18.31	16.38	19.31	11.88	13.31
	S.D.	1.31	3.24	5.98	4.87	5.76	4.09	6.85	7.10

* Indicates maximum possible score

THE JOURNAL OF EXPERIMENTAL EDUCATION

are clues in the following questions which will help you to work out the correct method.

4. You will not be able to ask any questions when you start again. Have you any questions now?"

The vast majority of Ss worked consistently, perhaps due to the novelty of the program and/or the constant pace demanded by the tape recorder. At the end of the session, E thanked them for doing so well, and asked them not to discuss the work in the booklets with anyone. (Ss were not told that they would be tested later.)

Testing session. Testing was held exactly 4 weeks after the learning session, generally at the same time of day. Due to school programming difficulties, one of the nine elementary school and two of the ten high school classes were tested 1 day early, and one small high school class was tested 2 days early.

To save time and prevent possible disorganization, Ss' names were written in advance on posttests A and B. On a random basis within each class, Ss were assigned to the egrule or the ruleg relearning program (i. e. posttest B).

The following instructions were given:

There are three separate parts to today's work. The first is to see how well you remember the work you did last time I was here. The second part is similar to what you did last time, and you will be working through the booklets in the same way as last time; remember to put your answers on the separate answer sheet. The third part is like the first. At the start and end of each part, write down the time from the board (signify position). Put up your hand when you finish each part. Try each question, but if you can't answer a question, go on to the next question; don't spend too long on any one question. Do the best you can. Commence now.

As each S finished one of the posttests, E collected it and gave him the next posttest; this was to ensure that the rules forming part of the relearning program (posttest B) could not be referred to by S when doing either posttest A or C.

RESULTS

The set of results is shown in Table 4. Analysis of variance computations were undertaken. Tests for homogeneity of variance were undertaken (Bartlett's method); where appropriate, transformations were

TABLE 5

VALUES OF F (ABOVE 1)

Factor	Task	A Retention	A Transfer Far	A Transfer Near	B Relearning Correct	B time	C Retention	C Transfer Far	C Transfer Near
Method (M)	C		1.4			2.2			2.0
	P	2.0		1.5			1.7	1.3	
Intelligence (I)	C	25.8**	6.7*	15.8**	33.8**		31.3**	32.1**	27.5**
	P	9.8**	20.9**	11.6**	25.8**	4.1*	15.9**	43.5**	17.9**
Grade (G)	C	22.1**	37.1**	95.5**	36.5**	63.9**	46.3**	38.2**	60.5**
	P	30.2**	52.5**	27.8**	18.4**	88.7**	19.0**	56.7**	24.2**
MI	C					1.5	5.4*		
	P								
MG	C					2.9			2.1
	P						1.6		
IG	C	6.3*	2.7			8.7**	2.7		2.8
	P		1.9		6.3*	1.1	1.8		
MIG	C	8.0**	1.5					1.2	
	P			1.4					

C - Concept Learning Task (Matriculation)
P - Principle Learning Task (Map-Reading)
** - Significant at 0. 01 level
* - Significant at 0. 05 level

made to the data (e. g. transformation of $\sqrt{x + 0.5}$ recommended by Bartlett (5), and further analysis of variance determinations computed. It was found that there was no difference in level of significance on any effect before and after the transformations; the given values of F in Table 5 are from the analysis with the lowest F levels.

Grade and Intelligence. As expected, grade 9 Ss performed significantly better than grade 5 Ss, and high intelligence Ss performed significantly better than average intelligence Ss.

Instructional Method. There were no significant differences between egrule and ruleg methods. Based on the significant and nonsignificant interactions, there would seem to be a slight superiority of the ruleg method for all groups except the grade 9 high intelligence group. Considering the large number of measures employed, the two significant interactions involving instructional method could be due to chance.

Category of learning. Based on graphic plots, significant interactions between category of learning and instructional method were obtained on the two sets of far transfer measures. For posttests A and C, on the far transfer measures, the performance of the egrule group was superior to the performance of the ruleg group on the principle learning task, contrasted to the superiority of the ruleg group on the concept learning task.

Sequence of learning. The utilization of the relearning measures enabled analysis into the effectiveness of the various sequences of learning, namely, egrule-egrule, egrule-ruleg, ruleg-egrule and ruleg-ruleg. For the six analyses (from the three sections of posttest C for each task), not one main effect or interaction involving instructional method was significant.

DISCUSSION

The results of no significant differences between instructional methods, and insignificant superiority of the ruleg method, are similar to other recent findings (23, 24, 25).

The interaction effects involving instructional method with grade and/or intelligence were, with two exceptions, insignificant. These exceptions, together with an analysis of the complete set of results, suggest that the egrule method is relatively more suitable for older students of high intelligence. From a recent apparently well designed experiment, Tanner (23) reported similar results: no significant main effects were found with respect to method of instruction, while there was a tendency for the discovery method to be relatively more effective for higher intelligence Ss. The present results do not support a number of experimental findings (3, 9, 18) in which it has been found that a discovery method is relatively superior for average intelligence Ss. The present findings question the prediction from Ausubel's hypothesis of superior results from the ruleg method for high school students, and superior results from the egrule method for elementary school children.

Interpretation of the interactions for both far transfer measures between instructional method and category of learning is difficult. Certainly the finding does not confirm the prediction from Gagne's hypothesis concerning superior results from the egrule method for the concept learning task. This is possibly due to the differences in the tasks, especially the differences in difficulty level between the two tasks.

The reasons for the lack of significance would appear to lie principally in the treatments. The tasks were non-arbitrary. Subjects participated in the experiment only if they failed the preliminary screening test. For one task (map-reading), both instructional groups performed equally well on the learning check items, indicating that this known possible problem of different amounts of learning was negligible for this task. The tasks, especially the map-reading task, were not too difficult; neither would they seem to have been too easy. A higher degree of control than that of many previous experiments was established by using programmed instruction booklets with the same verbal content and organization, and by randomly assigning Ss to treatments within classes in order to control class differences in attitude, achievement, and previous experience. The time spent by both instructional groups was exactly the same. The measures employed would appear to have been capable of discriminating any significant difference.

Why do the present results fail to confirm the findings obtained by Gagne and Brown (8), Scandura (19), Kersh (12), Ray (16) and Rowlett (17), who reported the majority of results significantly favoring the discovery method? The Ss in the discovery group in the experiments conducted by Gagne and Brown (8), Scandura (19) and Kersh (12) spent more time learning the material than the Ss in the corresponding ruleg group. In addition Ss in Kersh's (12) group spent more time on the task between the learning and testing sessions. Ray's (16) discovery group spent more time in manual activity than the ruleg group on a micrometer-reading task described by Grote (9) as "manipulative" in nature. From replications of Rowlett's (17) experiment under different conditions by Rowlett (18) and by Suess (21), insignificant results have generally been obtained. The specification of the differences between the discovery and ruleg methods in most of these experiments has been inadequate. Grote, who worked at Illinois at about the same time as Ray and Rowlett stated that his experiment was an extension of those by Ray (16) and Rowlett (17). He had this to say concerning the three experiments:

A primary difficulty rests in the ambiguity of the nature of the methods used by the experimenters. One cannot say, with any degree of confidence, that the directed discovery method in the various studies is similar or dissimilar, or that it contains the features or characteristics ascribed to the method by each investigator (9:119).

A large proportion of the previous experiments, if not all, have confounding variables interacting with the instructional method, in addition to other methodological problems. The weight of evidence for the significant superiority of the discovery method is thus very limited.

It is possible that a factor in the experiment could have resulted in one or the other of the groups being placed at a disadvantage. First, with respect to the ruleg group, it is possible that if less time had been allowed, ruleg group members may still have performed as well, but the egrule group members may have performed less well. Second, there are a number of factors which could be advanced as being to the disadvantage of the egrule group:

(a) The possible depressing effect of the presentation of the rule to the inductive group has been noted. However, the rule was first presented not at the end but about 60 percent of the way through each unit; the possibility of retroactive inhibition would thus be expected to be minimal.

(b) Programmed instruction may possibly not be the most appropriate way of catalyzing discovery. In experiments in which programmed instruction booklets have been used for both instructional groups (e. g., 23) the result has generally been insignificant. Perhaps discovery methods are most effective when a teacher is providing encouragement. (The value of programmed instruction with respect to control is, of course, immense.)

(c) Perhaps there was insufficient time and/or number of examples for egrule groups (apart possibly from the high intelligence grade 9 Ss) to "discover" the rule.

(d) The lack of experience in discovery methods and/or the lack of specificity in the instructions may have adversely affected egrule group members.

(e) The set rate of presentation may have affected the egrule group differently than the ruleg group.

(f) As the tests present a structured discovery form of learning of the $R_{ng} A_{ng}$ category (i. e., rule not given, answer not given), the ruleg Ss would also have had some discovery learning and could possibly have "discovered" the rules on one of the tests.

Many further avenues of research would appear useful. A number of extensions to the present experiment are possible:

An investigation into the significant interactions on the category of learning variable to determine whether these results were related to the difficulty of the task, the organization of the material, or the category of learning.

An investigation to determine what degree of specificity of instructions is required to optimize the performance of discovery Ss.

An investigation to determine what pre-training, if any, is required to optimize the performance of discovery Ss.

An investigation using Ss from grade 3 or 4. Ausubel suggested that approximately 12 years was the upper age at which discovery methods are most appropriate; perhaps the age of inflection is 10 years, assuming Ausubel's thesis to be valid. Possibly mental age should also be considered.

Analysis of the responses of Ss in discovery groups could serve as a basis for developing hypotheses. If programmed instruction booklets were used, this could be done by omitting the answers to some of the items. Such data would help to determine whether the Ss had learned the rule before its first presentation, the type of strategy used, and those factors, such as perseveration, which appear relevant when Ss learn a wrong rule by a discovery method.

Assuming that some forms of discovery learning are effective, an attempt could be made to analyze possible mechanisms. For example, it is possible that factors associated with "incidental learning" could also be relevant to discovery learning.

Personality factors could be very relevant to learning by discovery. For example, what is the effect on people of different anxiety levels of being placed in an "unstructured" learning situation.

If a grand research strategy to investigate learning by discovery were to be undertaken, it is considered that sets of reference lesson units in various subject areas and for various difficulty levels should be developed; appropriate pre- and posttests should also be prepared and standardized.

For any research project relating to learning by discovery, it is considered that there should be an emphasis on specifying the operations involved in the instructional method, on attempting to improve the terminology, and on attempting to overcome the methodological problems.

REFERENCES

1. Ausubel, D. P., "Learning by Discovery: Rationale and Mystique," National Association of Secondary-School Principals Bulletin, 45: 18-58, 1961.

2. Bruner, J. S., Goodnow, J. J.; Austin, G. A., A Study of Thinking, Wiley, New York, 1956.

3. Corman, B. R., "The Effect of Varying Amounts and Kinds of Information as Guidance in Problem Solving," Psychological Monographs, 71: (whole no. 431), 1-21, 1957.

4. Cronbach, L. J., "The Logic of Experiments on Discovery," in Shulman, L. S.; Keislar, E. R. Learning by Discovery, Rand McNally, Chicago, 1966, pp. 76-92.

5. Edwards, A. L., Experimental Design in Psychological Research, Rinehart, New York, 1957.

6. Gagné, R. M., The Conditions of Learning, Holt, Rinehart, and Winston, New York, 1965.

7. Gagné, R. M., "Varieties of Learning and the Concept of Discovery," in Shulman, L. S.; Keislar, E. R., Learning by Discovery, Rand McNally, Chicago, 1966, pp. 135-150.

8. Gagné, R. M.; Brown, L. T., " Some Factors in the Programming of Conceptual L e a r n i n g," Journal of Experimental Psychology, 62:3 1 3 - 321, 1961.

9. Grote, C. N., " A Comparison of the Relative Effectiveness of Direct-detailed and Directed Discovery Methods of Teaching Selected Principles of Mechanics in the Area of Physics," unpublished EdD. thesis, University of Illinois, Urbana, 1960.

10. Hermann, G., " Learning by Discovery: A Critical Review of Studies," The Journal of Experimental Education, 38:58-72, 1969.

11. Hunt, E. B., Concept Learning, W i l e y, N e w York, 1962.

12. Kersh, B. Y., " The Adequacy of ' Meaning' as an Explanation for the Superiority of Learning by Independent Discovery," Journal of Educational Psychology, 49:282-292, 1958.

13. Kersh, B. Y., " The Motivating Effect of Learning by Directed Discovery," Journal of Educational Psychology," 53:65-71, 1962.

14. La Rocque, G. E., " The Effectiveness of the Inductive and Deductive Methods of Teaching Figurative Language to Eighth Grade Students," Dissertation Abstracts, 26:6555A, 1966.

15. Leith, G. O. M., A Handbook of Programmed Instruction, University of Birmingham, Birmingham, England, 1966.

16. Ray, W. E., " Pupil Discovery Versus Direct Instruction," The Journal of Experimental Education, 29: 271-280, 1961.

17. Rowlett, J. D., " An Experimental Comparison of Direct-detailed and Directed Discovery Methods of Teaching Orthographic Projection Principles and Skills," unpublished EdD. thesis, University of Illinois, Urbana, 1960.

18. Rowlett, J. D., " An Experimental Comparison of Direct-detailed and Directed-discovery Methods of Presenting Tape-recorded Instruction," in Rowlett, J. D., Status of Research in Industrial Arts, 15th Yearbook of the American Council on Industrial Arts Teacher Education, Illinois, 1966, pp. 123-155.

19. Scandura, J. M., " An Analysis of Exposition and Discovery Modes of Problem Solving Instruction," The Journal of Experimental Education, 33: 149-159, 1964.

20. Silberman, H. F., " Characteristics of S o m e Recent Studies of Instructional Methods," i n Coulson, J. E., Programmed Learning and Computer-based Instruction, Wiley, New York, 1962, pp. 13-24.

21. Suess, A. R., " The Effect of Manipulation o n the Directed Discovery Method of Teaching Orthographic Projection Principles," 1965, quoted in Householder, D. L., " Techniques and Modes of Instruction," Review of Educational Research, 38: 382-394, 1968.

22. Tanner, R. T., " Discovery as an Object of Research," School Science and Mathematics, 68: 647-655, 1969.

23. Tanner, R. T., " Expository-deductive Versus Discovery-inductive Programming of Physical Science Principles," Journal of Research in Science Teaching, 6:136-142, 1969.

24. Werdelin, I., " The Value of External Direction and Individual Discovery in Learning Situations: I. The Learning of a Mathematical Principle," Scandanavian Journal of Psychology, 9: 241-247, 1968.

25. Werdelin, I., " The Value of External Direction and Individual Discovery in Learning Situations: II. The Learning of a Foreign Alphabet," Scandanavian Journal of Psychology, 9:248-251, 1968.

26. Wittrock, M. C.; " The Learning by Discovery Hypothesis," in Shulman, L. S.; Keislar, E. R., Learning by Discovery, Rand Mc Nally, Chicago, 1966, pp. 33-75.

APPENDIX

EXAMPLES OF MAP-READING MATERIALS

(a) Learning Program

   Items 24-31

   Map 2 (Appears at the top of the next page).

Corresponding egrule example	Ruleg item	Corresponding egrule item *	Time (secs) Grade 9	Grade 5
(E24) Long Bay	(R24) On Map 2, we will use the horizontal (or sideways) lines to find the northing of some suburbs. On Map 2, the position of Long Bay, with respect to the horizontal lines, is found by looking from bottom to top (from south to north) and by finding the third digit using the scale. The northing of Long Bay is_____.	E28	50	75
(25) Bondi	(25) Carefully look at the scale on the map to see how the scale is used to find the northing of Bondi. The northing of Bondi is____	25	30	45

MAP 2

(Smaller than actual size.)
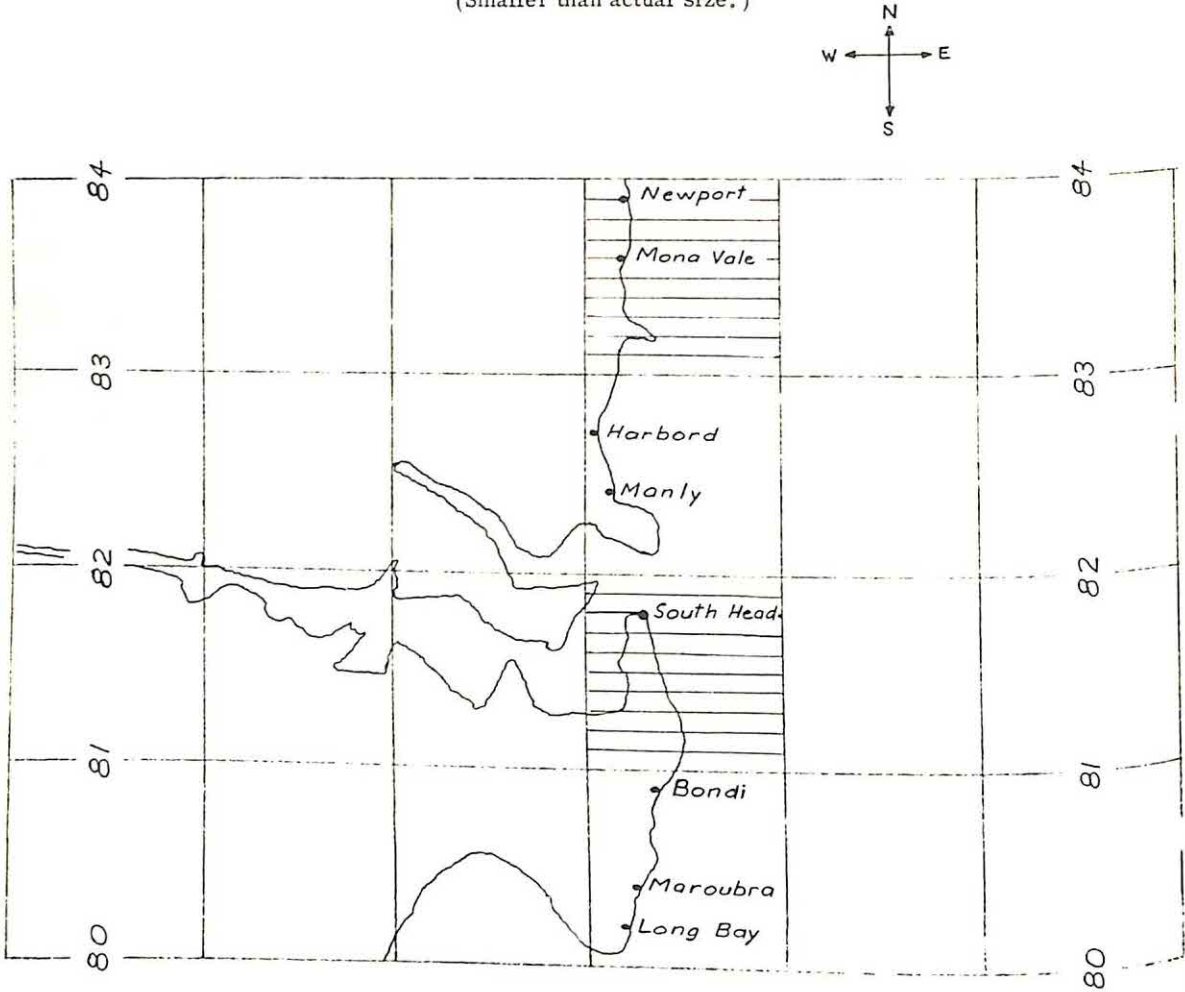
(E26) New- port	(R26) For Newport, the northing is found by count- ing 9 units upwards (to- wards the top) from the 83 horizontal line. Check on Map 2 using the scale. The northing of Newport is ___.	E30	30	45
(E27) Harbord	(R27) Remember to look from bottom to top. The northing of Harbord is ___.	E31	20	30
(E28) Maroubra	(R28) On Map 2, the hor- izontal (or sideways) lines are used to find the northing of some suburbs. The northing of Maroubra is ___.	E24	50	75

(29) Mona Vale	(29) In a map reference, there are 6 digits; three dig- its refer to the easting and three digits refer to the north- ing. The northing of Mona Vale is ___.	29	30	45
(E30) South Head	(R30) The northing of South Head is ___.	E26	30	45
(E31) Manly	(R31) The northing of Manly is ___.	E27	20	30

* Except for the specific example used.

(b)   Examples of Far Transfer test items.

MAP D (Smaller than actual size.)
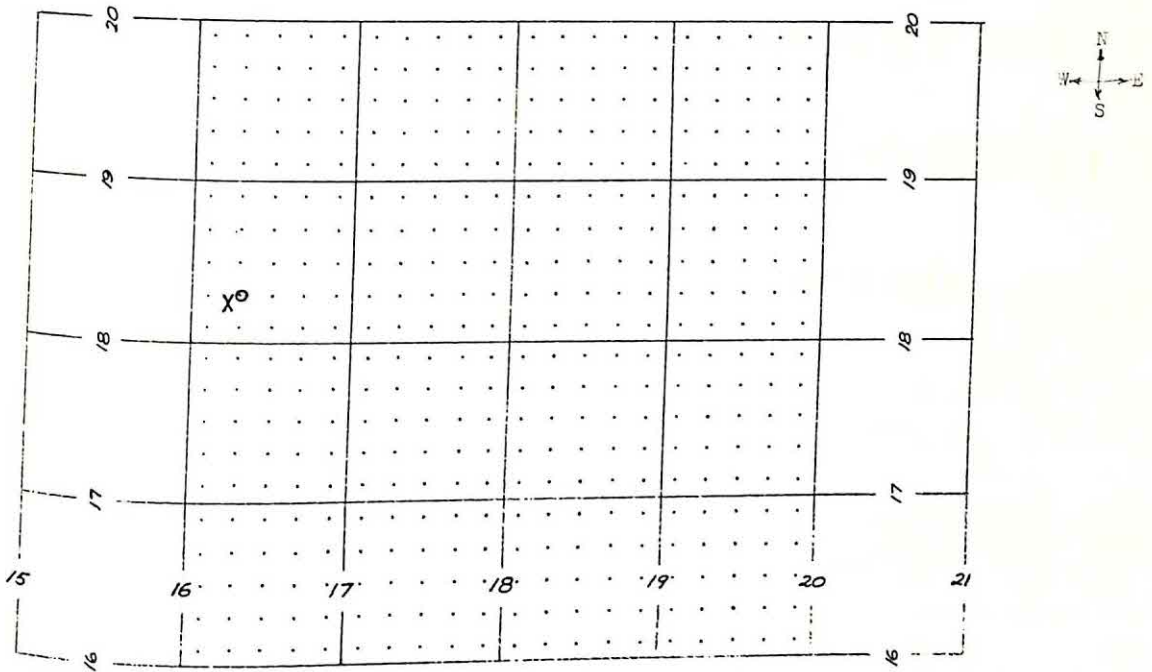
Note how position **X** is shown. Similarly, put in position R at 193189 <u>on the map</u>.

MAP E (Smaller than actual size.)
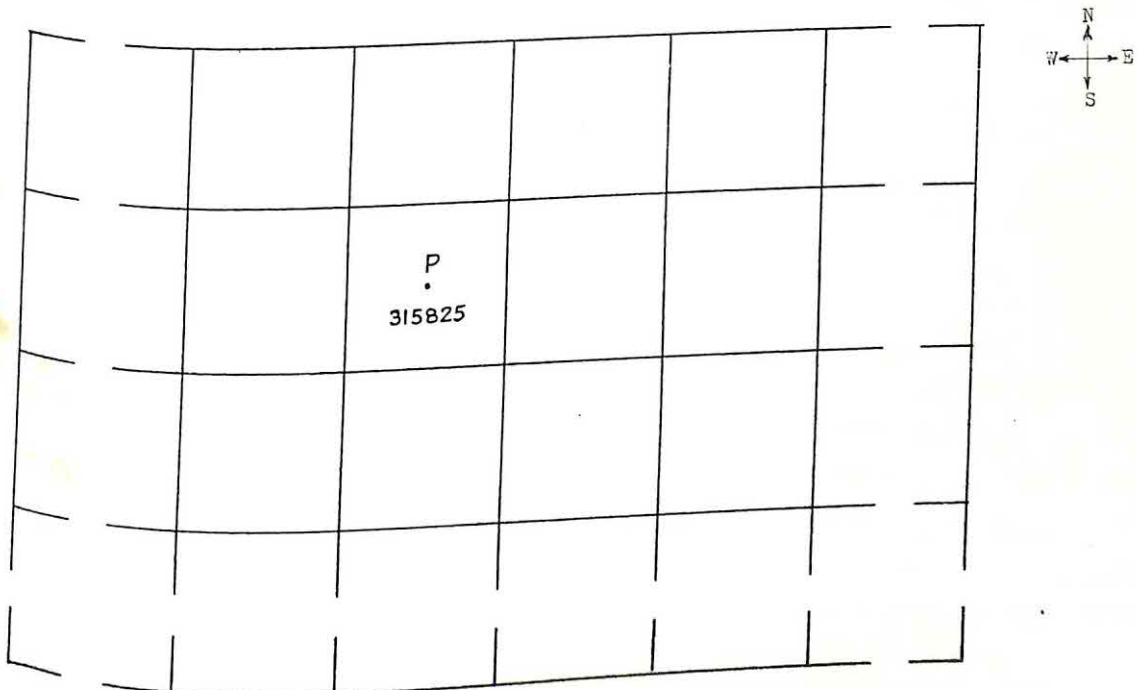
Using the position of **P** (315825), write <u>on the map</u> the missing numbers of the horizontal and vertical lines.

THE TEACHING ASSESSMENT BLANK:

A FORM FOR THE STUDENT ASSESSMENT

OF COLLEGE INSTRUCTORS[1]

DAVID S. HOLMES
University of Texas, Austin

ABSTRACT

A form which students can use to assess their class experiences was presented. A factor analysis based on evaluations filled out by 1,648 students revealed four factors which measured (a) the quality of the instructors' presentations, (b) the evaluation process and the student-instructor interactions, (c) the degree to which the students were stimulated and motivated by the instructors, and (d) the clarity of the tests. A further analysis indicated that subscale scores which reflected the factor scores could be developed from the total item pool.

THERE IS a long and varied history to the systematic evaluation of college instructors by their students. Because of the increasing external and internal demands being placed on universities for better teaching, increasing student concern for the quality of instruction, and the fact that it is becoming increasingly difficult or impossible for administrators to visit and evaluate all classes, there is good reason to believe that in the future the evaluation of instructors by students will be more widespread and relied on more heavily.

The Teaching Assessment Blank (TAB)[2] was designed and is being used by the University of Texas at Austin with three goals in mind: (a) to provide data for a publication to be used by students interested in obtaining information about prospective instructors (2); (b) to provide the university administration with data upon which to base, in part; their evaluation of members of the teaching faculty; and (c) to provide the faculty members with feedback in hopes of improving their teaching. It is important to note that the scale was not designed to evaluate an instructor's academic competence or knowledge or the value of the course he teaches. Instead the items on the scale are limited to those related to the instructor's teaching ability. The TAB was constructed with this limitation since it could be argued that many students are not in a position to evaluate an instructor's academic competence or the value of his course. On the other hand, college students are in a position to evaluate the way in which the instructor presents his material. Furthermore, students are in a better position than anyone else to report their own responses to the instructor's presentation, i.e., the interest, effort, and thought provoked in them by the instructor. The ability to stimulate students would seem to be an important aspect of teaching to evaluate, especially if learning is considered to be an active process. The TAB is unique in that unlike previous

evaluation forms it has a number of items which directly assess this aspect or result of teaching.

In its present form (Form II), the TAB has thirty-five statement-items (see Table 1), some of which elicit factual responses ("My overall grade average is . . ."), while others elicit subjective responses ("The instructor seemed to be well-prepared for lecture or discussion.")[3] The items are printed on an answer sheet designed for electrical or mechanical scoring. The TAB is filled out anonymously at the end of the course. The instructions are:

Please evaluate this instructor by indicating the one response that most nearly describes the feeling you have had generally or most of the time concerning his teaching. Your evaluations will be of great value: (1) if you state your own personal feeling without considering what the sentiment of others might be; and (2) if you realize that, while it is difficult to respond exactly to the kinds of statements contained below, with care you can select one response. Use a Number 2 or 2 1/2 pencil. Omit only those items which are not applicable, or which you do not feel qualified to judge.

The items from the TAB are presented in Table 1. On the basis of their content, these items can be broken down into three types. The first group contains six items which provide basic information about the respondent: year in school (item 1), sex (item 2), grade point average (GPA) (item 4), whether the course was in his major area (item 5), college membership (item 34), and reason for taking the course (item 35). These were included primarily so that persons considering the course or the published evaluation could identify the characteristics of the students previously enrolled in it. The second group contains twenty-four items related to specific

TABLE 1

ITEM STEMS FROM THE TEACHING ASSESS-
MENT BLANK (FORM 2)*

1. My classification is:
2. Sex:
3. My final grade in this course was, or probably will be:
4. My overall grade average at U. T. is:
5. This course is part of my major field:
6. I wanted to take this course before the semester began:
7. The instructor seemed to be well-prepared for lecture or discussion:
8. He used enough examples and illustrations to clarify the materials for me:
9. He presented material in a coherent manner, emphasizing major points and making clear their relationships:
10. He usually was aware of whether the class members were following his discussion or lecture with understanding:
11. He usually held my attention during class:
12. The instructor, through the course, challenged me to be creative in my work:
13. He made me feel free to ask questions, disagree, and express my ideas:
14. He was fair and impartial in his dealings with students:
15. He was intellectually stimulating (he caused me to think):
16. He revealed enthusiasm for his teaching:
17. He let us know what he expected of us on tests and assignments:
18. The meanings of questions on his tests were usually clear:
19. He usually returned assignments and tests promptly:
20. He had sufficient evidence, in terms of class participation and written work, to evaluate the quality of my performance in this course:
21. His grading system was fair to me as an individual:
22. He was readily available for conference outside the class:
23. He seemed to be interested in students as persons:
24. He interested me in the subject of this course:
25. I looked forward to attending class:
26. In comparison with all the instructors I have had, both in high school and college, this instructor was:
27. In comparison with all the courses I have had, both in high school and college, this course was:
28. All things considered, the text book used in this course was:
29. I made an honest effort to learn in this course:
30. In this course I learned a great deal:
31. The instructor made clear to me his educational objectives in this course:
32. The instructor accomplished his educational objectives in this course:
33. I tried to meet with the instructor outside of class:
34. The college or school in which I am now enrolled is:
35. I took this course to satisfy:

*Response alternatives for items 6-25, 29-32: Definitely Yes, Yes, No, Definitely No; for items 26 and 27: One of the best, Above average, Average, Below average, Far below average; for item 28: Excellent, Above average, Average, Below average, Poor; for item 33: Many times, A few times, Only once or twice, Never; for item 34: Arts and Sciences, Business Administration, Education, Engineering, Other; and for item 35: Major or minor field requirements, Other specific degree requirements, Elective credits required for degree, Non-degree requirements (e. g., requirements for teacher certification), No requirements at all.

aspects of the instructor and course under consideration; that is, items which assess the instructor's presentation, examinations, etc. The items in this group will be discussed and classified further with regard to the factor analysis presented in a later part of this paper. The third group contains five items with miscellaneous content; two (items 22 and 33) deal with the out of class contact with the instructor, one (item 28) inquires about the quality of the textbook used in the course, one (item 5) asks about the grade expected in the course, and one (item 6) assesses the degree to which the student wanted to take the course before enrolling.

STUDY I: FACTOR ANALYSIS OF CLASS RELEVANT ITEMS

SUBJECTS

During the spring semester of the 1967-68 academic year, faculty members of the College of Arts and Sciences at the University of Texas were invited to participate in the class-instructor evaluation program in which the TAB (Form II) was the measuring instrument. Evaluations were carried out in 322 classes. From these, seven large classes with a total of 1,648 students responding were selected on a roughly random basis to provide the data for the present analyses. Involved were three classes in history, two in geology, one in zoology, and one in psychology. The classes ranged in size from 106 to 358. Overall, there were 480 freshmen, 586 sophomores, 394 juniors, 166 seniors, and twenty-two graduate students. The use of only large classes may limit somewhat the generalizability of the results to be reported since there may be factors which influence the evaluation responses in small classes which may not be operating in large classes. Large classes were specifically chosen for study because the success of these classes is more dependent upon the quality of the instructor's presentations than it is in small classes and, since in the future this type of class will no doubt play a dominant role in undergraduate's university experience, it seemed important to focus attention on this situation. Research to be reported subsequently will deal primarily with the small class.

ANALYSIS

The factor analysis was a principal axis analysis with a varimax rotation (3). The analysis was carried out on twenty-two of the twenty-four class-related evaluation items. Items 26 (comparison with all other instructors) and 27 (comparison with all courses) were not included since they involved

TABLE 2

VARIMAX LOADINGS FOR EVALUATION
RELEVANT ITEMS

Items (arranged by largest factor loading)	FI	FII	FIII	FIV
Factor I: INSTRUCTOR PRESENTATION				
7. instructor well prepared	.80	.10	.08	.19
16. revealed enthusiasm for teaching	.77	.24	.15	.09
9. presented material in coherent manner	.76	.05	.21	.33
8. enough examples and illustrations	.75	.06	.17	.34
10. aware of class following discussion	.58	.34	.25	.19
Mean Loadings	.73	.16	.17	.23
Factor II: INTERACTION-EVALUATION				
19. returned assignments-tests promptly	.02	.72	.07	.05
13. feel free to ask questions, disagree	.39	.63	.12	.10
20. sufficient evidence to evaluate	.01	.61	.14	.50
21. grading system was fair to me	.06	.56	.16	.56
23. interested in students as persons	.48	.52	.23	.15
14. fair and impartial with students	.32	.48	.18	.25
Mean Loadings	.21	.59	.15	.27
Factor III: STUDENT STIMULATION				
30. I learned a great deal	.20	.13	.77	.21
29. I made honest effort to learn	.13	.11	.74	.06
25. I looked forward to attending class	.52	.14	.63	.06
24. he interested me in subject	.58	.16	.59	.15
15. stimulating . . . caused me to think	.52	.21	.58	.13
11. held my attention during class	.55	.07	.55	.13
12. challenged me to be creative	.42	.30	.54	.41
32. instructor accomplished objectives	.40	.17	.52	.13
31. instructor made clear his objectives	.36	.22	.48	.38
Mean Loadings	.41	.17	.60	.19
Factor IV: TEST CLARITY				
18. test questions clear	.36	.08	.21	.71
17. let us know what was expected on tests	.48	.10	.17	.63
Mean Loadings	.42	.09	.19	.67

summaries rather than measurements of specific
characteristics. The rotated factor loadings f o r
each of the items are presented in Table 2 where the
items are arranged by their highest factor loadings.
A discussion of the contents and meanings of the fac-
tors follows.

Factor I

An inspection of the content of the items showing
highest loading on this factor clearly indicates that
it constitutes an evaluation of the instructor's pre-
sentation. The items measure, for example, the in-
structor's preparedness, enthusiasm, organization,
use of examples, and awareness of students' under-
standing. It is important to note that the items with
highest loadings on Factor I showed relatively l o w
loadings on the other three factors. The question of
the independence of the factors will be discussed in
greater detail later. In view of the content of t h e
items with high loadings on this factor it will be re-
ferred to as the Instructor's Presentation f a c t o r.
Factor I accounts for 23.97 percent of the variance.

Factor II

Factor II seems to be measuring two somewhat
different but highly related issues; first the instruc-
tor's system for evaluating the performance of the
students (grading system fair, sufficient evidence
to evaluate, returned assignments promptly) while
the second was the instructor's relations with stu-
dents (interested in students as persons, free to ask
questions, fair, and impartial). At first these might
seem to be quite separate but it could be suggested
that in classes with literally hundreds of students the
assignments, examinations, and grades provide the
major medium of interaction with the instructor for
most students; for this reason the issues are fused.
That is, the student perceives the instructor's atten-
tion to the evaluation process to be the only personal
response possible. In view of the content of the items
with the highest loadings on Factor II, this factor will
be referred to as the Evaluation-Interaction factor.
This factor accounted for 11.73 percent of the vari-
ance (third in order of magnitude) and, as was the
case with Instructor's Presentation factor, the items
which have their highest loadings on this factor have
relatively low loadings on the other factors.

Factor III

The items with highest loadings on Factor III are,
for the most part, those which measure student re-
actions and behaviors which are evoked by instruc-
tors. That is, in contrast to the items making up
Factors I and II which directly assessed the actions
of the instructor (his presentation a n d evaluation
system), the items with highest loadings on Factor
III assessed what the students did in response to the
instructor's behavior. For example, the items mea-
sured the degree to which as a function of the instruc-
tor the students felt they learned, made an effort,
looked forward to attending, became interested, were
stimulated, thought, and were challenged to be cre-
ative. It is important to note that the behavior as-
sessed by these items were responses—responses
to the actions of the instructor—and therefore these
items provide interesting additional, indirect data on
the instructor's presentation, specifically, its effect.
Due to the content of the items loading heaviest on
this factor, it was called the Student Stimulation fac-
tor. This factor accounted for 16.82 percent of the
variance (second in order of magnitude).

It should be noted that there are two items which
have highest loading on this factor but which are in-
consistent with the other seven items in t e r m s of

content. These items deal with the clarity of course objectives (item 31) and the attainment of the objectives (item 32). The relation of these items to the factor in general is not clear, though it may be that clear goals and their attainment facilitates enthusiasm. Of the items with a primary loading on this factor, items 31 and 32 have the lowest loadings.

Factor III does not seem to be as independent as Factors I and II; the items which make up Factor III also tend to load on Factor I. While the items with the highest loadings on Factor I show a considerable difference in mean loadings on Factors I and III (.73 versus .17), the items with the highest loadings on Factor III show considerably less difference in mean loadings on Factors I and III (.41 versus .60). It seems that a high rating in terms of presentation (Factor I: preparedness, enthusiasm, coherence, etc.) can be relatively independent of the degree to which students are stimulated (Factor III: interest, effort, etc.). On the other hand, if the students are stimulated (Factor III), the instructor is likely to be judged as having given a good presentation. To use theatrical terms, an instructor may give a "technically" good performance but may not "project across the footlights." To project across the footlights, however, usually requires a good performance. It seems then that a technically good presentation is helpful, but not always sufficient, to motivate students in large lecture classes. While the present data did not directly indicate what in addition to the presentation caused students to become motivated, the degree to which students are stimulated would seem to be an important factor to measure in evaluating instructors, especially if the goal of teaching is seen as more than conveying information, a process which a textbook may do more effectively and less expensively.

Factor IV

Only two items have highest loadings on Factor IV but their content is quite consistent. Both items seem to be measuring the clarity of the instructor's tests, therefore, this factor will be called the Test Clarity factor. These items have second highest loadings on the Instructor's Presentation factor, a finding which suggests that the preparedness, coherence, and awareness of students' understandings which facilitated the presentation of material tended to be reflected also in the preparation of examinations. Factor IV accounts for 10.47 percent of the variance.

In view of the preceding findings and discussion, it can be concluded that the class relevant items of the TAB fell into four both statistical and logical subgroups and that each subgroup measured a different, and in most cases, relatively independent aspect of college teaching. It might be noted that the above factor structure was essentially replicated with another selection of large classes, thus suggesting that the factor structure is rather stable.

STUDY II: DEVELOPMENT OF SUBSCALES

In an effort to simplify the interpretation of the TAB, an attempt was made to arrive at a system of determining summary scores for each of the four factorially identified areas, thus reducing the number of scores from twenty-four to four. One approach

to this problem is to use factor scores derived from a factor analysis like that described above. While statistically elegant, this approach requires that a factor analysis be carried out each time factor scores are wanted and this is impractical or impossible for most users of the TAB. A second approach is to use the factor loadings from the above analysis to identify items for membership in subscales from which sub-scores could be easily derived. That is, items with the highest loadings on a factor could be used to construct a subscale to represent that factor. Subscale scores would not be perfect representations of the factors since in the computation of subscale scores any item would contribute all of its variance to the one subscale to which it belonged; while in the computation of factor scores an item can simultaneously contribute different amounts of variance to different factors. The resulting discrepancy may not be important, however. The degree to which the subscales actually reflected the factor structure as revealed by the factor analysis can be determined by comparing the subscale scores with the actual factor scores. Through the use of a multi-trait (four aspects of teaching), multi-method (two types of scores)[4] matrix (1) it can be determined whether the scores of the subscales were similar to the factor scores they were designed to represent (convergent validity) and whether, as one would hope, these relationships were higher than those between the various subscale scores (discriminant validity).

Item selection

A subscale was constructed to represent each factor previously identified. The sole determinant of an item's subscale membership was its loading on the factor the subscale was designed to measure and therefore subscale item membership was identical to the factor item membership outlined in Table 2.

Multi-trait—multi-method matrix

Weighted factor scores for Factors I, II, III, and IV were determined for each S using the data from the factor analysis presented earlier. Subscale scores for subscales one, two, three, and four were determined for each S by summing the responses given to the items within each subscale. The intercorrelations between these right scores are presented in Table 3 in the form of a multi-trait—multi-method matrix.

Of primary interest are the values in the validity diagonal (underlined in Table 3) and the values in the hetero-trait—mono-method triangle for subscale scores (enclosed by a solid line in the lower right of Table 3). It is clear from the magnitude of the validity values (i.e., the correlations between the factor scores and the subscale scores) that the simple subscale scores reflected the factor scores rather well; that is, there seems to be a considerable amount of "convergent validity." In fact, the mean validity value is .71, thus indicating that on the average the subscale scores account for 50 percent of the variance in the factor scores. As would be hoped, the hetero-trait—mono-method values for the subscale scores (i.e., the correlations between the various subscale scores) are considerably lower than the validity values. That is, there is some degree of "discriminant validity." The mean is .51, thus indicating that on the average the score on one scale

TABLE 3

MULTI-TRAIT—MULTI-METHOD MATRIX FOR FOUR ASPECTS OF TEACHING, (1) INSTRUCTOR'S PRESENTATION, (2) EVALUATION-INTERACTION, (3) STUDENT'S RESPONSE, (4) TEST CLARITY, MEASURED BY TWO SCORES*

*The validity values are underlined. Each hetero-trait—monomethod triangle is enclosed by a solid line. Each hetero-trait—hetero-method triangle is enclosed by a broken line.

accounts for only 25 percent of the variance of the score on another scale. Ideally of course the hetero-trait—mono-method correlations could be lower but it should be noted that some degree of relationship would be expected between the subscales, especially, for example, between subscales one and three which measure the quality of an instructor's presentation and the degree to which students are stimulated. (The correlation between these two subscales is actually as high as one of the validity correlations.) It does appear, however, that there is enough separation between the scales to warrant their use. The use of the subscale scores will provide the user with more specific summary information than has been available in the past from one global score.

DISCUSSION

In view of the preceding analyses, it is clear that the TAB is a potentially very useful instrument for carrying out student evaluations of college instructors. It is evident from the factor analysis that the TAB measures a number of different but very relevant aspects of teaching. The existence of this multidimensionality and the fact that the various dimensions can be easily scored makes it possible to obtain summary evaluations on a number of areas within the broad topic of "teaching competence." In addition to being helpful in terms of providing feedback to the teacher, the specificity offered by the existence and identification of the subscales will also be of value to the researcher who all too often in the past could only work with a global score which reflected the entire constellation of teaching competence. This refinement of the measuring instrument may enable the researcher to more accurately describe the influence of other variables on teaching or teacher evaluation.

Of particular interest and potential value is the identification of the student stimulation factor. The degree to which instructors instill interest, enthusiasm, and motivation seems to be an important aspect to measure if we see the classroom as only the beginning of education and our goal as the turning of "pupils" into "students."

FOOTNOTES

1. The present research was carried out during the author's tenure as Research Consultant to the Measurement and Evaluation Center, University of Texas, Austin.

2. The credit for the initial development of the scale discussed in this paper belongs to Dr. Paul Kelly, Director of the Testing and Evaluation Center, and Miss Caroline Dowell, Research Associate, University of Texas.

3. In addition to responding to the thirty-five objective items, the students are invited to make free response comments on the back of the answer sheets. These are not considered in the present analysis.

4. It should be noted that in this analysis the different "methods" were not different methods of collecting the data but rather different methods of determining the scores.

REFERENCES

1. Campbell, D.; Fiske, D., "Convergent and Discriminant Validation by the Multitrait—multimethod Matrix," Psychological Bulletin, 56: 81-105, 1959.

2. Riddell, J., Student Guide to Courses and Instructors at the University of Texas at Austin, University of Texas, Austin, 1968.

3. Veldman, D., Fortran Programing for the Behavioral Sciences, Holt, Rinehart, and Winston, New York, 1967.

STUDENT ORIENTATION TO

AN INDIVIDUALIZED EDUCATION SYSTEM

G. BRIAN JONES, DANIEL W. KRATOCHVIL,

DENNIS E. NELSON, and WILLIAM E. STILWELL[2]

ABSTRACT

This was an initial attempt to examine experimentally the orientation of 165 fifth graders and 114 ninth graders to an individualized educational system. Little support was found for the hypothesis that the amount of orientation would have an effect on students' academic performance, and their opinions and knowledge of the system. The following guidelines, based on the interpretations of the results of this study, were suggested for future research in this area: (1) include pre- and post-measures of student orientation needs on each criterion; (2) employ a no-orientation control group; (3) control for the effects of inter-treatment modeling by students; (4) administer criterion measures related to student orientation needs and the orientation instructional objectives; and (5) include relevant indices of teacher ability to implement orientation programs.

DURING THE initial year (1967-68) of individualized education in Project PLAN (2) the observations of teachers, administrators, and field consultants suggested that students needed experiences designed to orient them to this individualized educational system. In this framework, appropriate student responses varied markedly from those in more conventional contexts. Extensive data, not too systematically collected, suggested that the longer a student had experienced conventional instruction, the more difficult it was for him to make the transition to Project PLAN. Staff observations indicated that specific orientation activities might reduce the number of students who appeared to flounder academically during the first months of the school year.

While this information seemed to emphasize the need for student orientation activities in Project PLAN, little consistent research support or direction for the development of such learning experiences was available. Four problems were involved. First, in no studies were student orientation needs systematically assessed before the implementation of various orientation programs. Even subjective data like those documented in Project PLAN were not available in the literature. Most investigators (8, 9) assumed the existence of student orientation needs including those labeled academic-

intellectual, social and informational.

Secondly, since orientation research was undertaken predominantly at the college level, it was difficult to generalize implications for the elementary and secondary students of interest in this study. Thirdly, too few studies assessed the effects of two or more orientation procedures in order to determine the best means of meeting students needs. Packard (14) provided an example of one of these studies. Using random assignment of students and criteria including a college information test, an attitude measure employing Osgood's Semantic Differential Technique, and a registration advisor's rating of each student's preparedness for the registration process, he found that a lecture approach was more effective in orienting students than were programmed workbook strategies. However, it appears that his results were not replicated.

Finally, evidence indicating whether or not students' needs were met by orientation programs was conflicting. The findings of several studies (4, 5, 10, 11, 12, 13, 14, 15) suggested, on the basis of positive student and staff reactions, that orientation was worthwhile. Yet the results of other similar studies using more varied criteria indicated that orientation may have been a waste of student and teacher time (1, 6, 16) or might have even hindered

student performance (3).

Assuming the validity of the available subjective data on Project PLAN students' orientation needs, the present study focused on the last three problems. Its Ss were fifth and ninth graders, not college students. It employed a design enabling a comparison of two orientation procedures and using more rigorous and extensive criteria than just student and staff reactions. The effects of the amount of orientation on student knowledge, opinions, and overt behavior including academic performance were examined. Criterion measures were developed to relate closely to Project PLAN student orientation objectives and this contrasted with many previous orientation studies in which instruments, apparently only indirectly related to the programs' orientation objectives, were employed. It was hypothesized that students who participated in a comprehensive orientation program would perform better academically, have more knowledge of the educational system, and have more favorable opinions about the system than would students who participated in a brief orientation program where the amount of information, student practice of relevant behaviors, and involvement provided was reduced to that deemed the essential minimum.

METHOD

Subjects

The sample included 114 ninth-grade students from two schools and 165 fifth-grade students from four schools. All students had been selected at random for Project PLAN's individualized education program. All schools were in the San Francisco Bay Area.

Orientation Programs

Student orientation materials consisted of booklets containing orientation information and suggestions as to what students might do and use to achieve the instructional objectives of the orientation program.

The orientation objectives were sequenced so that students proceeded from an introduction to the new education program, to learning the specific behaviors needed to function in the system, and on through individual planning and scheduling of their work in each subject matter area. The variety of learning activities recommended included reading, teacher or student demonstration of instructional equipment and materials, group discussions, individual planning sessions with teachers, and practice sessions for the performance of skills important in the individualized education system.

Two versions of these orientation materials were prepared for each of two grade levels (i. e., fifth and ninth). One version constituted a comprehensive and the other a brief orientation program. The versions differed in two major respects. The amount of information and behavioral practice included in the brief version was substantially less than that included in the comprehensive version. For example, students in the brief orientation program had fewer opportunities to practice operating audio-visual equipment and to participate in group discussion of various aspects of the educational system. The brief version also

allowed each student less involvement in planning and decision making relative to his academic work. For example, the teachers of the students in the brief program specified the type and amount of work each student would try to do; while the teachers of the students in the comprehensive program met with each student to discuss the type and amount of work each student would try to achieve. The comprehensive orientation program consumed, on the average, about 5 days of student time, while students completed the brief program in 2 days or less.

Criterion Instruments and Measures

As noted earlier, no systematic, objective data on student orientation needs were available on which to base the development of this study's criterion instruments. Reliance was placed upon previous observations of students by school personnel and upon a study of the materials and procedures with which students in the system would have to cope if they were to perform successfully. Therefore, the validity of instruments used in this investigation is limited to face validity.

Instrument development was based on the specific instructional objectives of the orientation materials, and the outcomes of student knowledge, attitude, or overt behaviors, which indicated the achievement of these objectives. Thus, all parts of each of the following instruments were keyed to the orientation objectives.

Orientation Module Tests Number 1 and Number 2. To test Ss' knowledge of the PLAN program immediately after they completed their orientation activities, two multiple-choice tests were written at each grade level. These tests were prepared in the same format as the subject-matter tests students regularly took in this individualized program. One test (21 items for fifth graders and 48 items for ninth graders) sampled Ss' knowledge and performance relevant to the first two orientation objectives, while the other test (10 items for fifth graders and 10 items for ninth graders) focused on a second pair of orientation objectives. Representative examples of an instructional objective, a related student outcome, and a test item for fifth graders follow.

Sample Instructional Objective:

To describe the responsibility and role for PLAN student.

Representative related student outcome:
Know and be able to perform procedures for working through teaching-learning units and for taking end-of-unit tests.

Sample Test Item:

If you receive an NP on a module test, your next learning activity should be to

A. stop working on Project PLAN for a while.
B. go on to the next module.
C. take another TLU in the same module, or repeat the same TLU.
D. tell your friends.

Orientation Survey Test. To check Ss' retention

and subsequent acquisition of knowledge, a 54-item multiple choice test was administered at the fifth-grade level 3 weeks after the orientation program. Similarly, a 65-item test was simultaneously given to ninth-grade Ss. Both tests sampled the same objectives and related knowledge outcomes assessed by the Orientation Module Tests, including knowledge of how the PLAN system functions, terminology and procedures used, and responsibilities of students and school personnel. A sample question from the secondary level survey test follows.

Sample Test Item:

The two ** signs in front of objective 3186 mean that this student should
A. know he did exceptionally well on that objective.
B. know that this objective has a double number.
C. review this objective and then meet with his teacher.
D. take the test for that objective again.

Structured Interviews. Two weeks after orientation, 20-minute interviews were administered by research assistants using structured questionnaires. These interviews assessed the extent to which randomly selected Ss in each treatment were able to express their understandings of selected concepts and procedures in PLAN and to perform behaviors which were necessary for Ss to function in PLAN classrooms.

Once again, an example of an instructional objective is provided below along with a representative student outcome and an example test item. All interview items were structured to permit straight-forward scoring as either "pass" or "fail" by the interviewers using detailed sample scoring keys.

Sample Instructional Objective:

To find and to operate correctly all learning aids in a PLAN classroom.
Representative related student outcome:
Correctly operate tape recorder and tapes.
Sample Interview Item:
Ask the student to go to the tape recorder, put on a tape, and begin playing it.

Opinion Surveys. As measures of students' expressed opinions toward the PLAN educational system, a 14-item survey for fifth-grade Ss and a 16-item form for ninth graders were developed. The possible effects of response set were counterbalanced by phrasing the statements so that to agree with some would represent a favorable opinion of PLAN while agreement with others would represent an unfavorable opinion of PLAN. Degree of agreement or disagreement on a 4-point scale (to force a decision in one direction or the other) was requested. Available responses were "strongly agree," "slightly agree," "slightly disagree," and "strongly disagree." The opinion tests were administered at the same time as the Orientation Survey Tests, 3 weeks after the orientation.

Sample Instructional Objective:

To describe the responsibility and role of a PLAN student.

Representative related student outcome:
Know and appreciate that a major decision maker in PLAN is the student.
Sample test item:
The teacher makes all the decisions in a PLAN class.

Academic Performance Records. Constantly maintained within Project PLAN are computer records of students' performance in their four subject matter areas-Language Arts, Mathematics, Social Studies, and Science. Students usually demonstrate their successful achievement of a group of instructional objectives—called a "module"—by taking a module test whenever they and their teachers agree that they are adequately prepared. Not all students work on the same modules; however, all modules are intended to take a roughly equivalent amount of instructional time for the average student for whom they are appropriate. A computer printout of all Ss' academic records was provided 10 weeks after the orientation activities. This enabled a frequency count of the number of learning units students had successfully completed during these 10 weeks.

PROCEDURES

Within each participating classroom, students were randomly assigned, regardless of sex, to either the comprehensive or the brief orientation program. It was impossible to implement an inactive control treatment because teachers refused to withhold all orientation assistance from some of their students. Such refusal might be ascribed to their observations of the deleterious effects of such a treatment during the previous year of Project PLAN. Similar negative reactions were received regarding possible pretreatment administration of criterion measures. Since there was a reluctance to greet students with a series of tests immediately upon their return to school in September, no pretest data were collected at that time. Therefore, this study's design had to rest upon previously gathered information on PLAN students' orientation needs, random assignment of Ss, and analysis of variance statistics on post-treatment data only.

With the help of research assistants and printed guidelines, each PLAN teacher administered the two treatment procedures without informing students that an experiment was being conducted. Because students in any PLAN classroom seldom work on the same materials, there were no problems in introducing two sets of orientation materials. All criterion instruments except the structured interviews were administered to all students. One half of the Ss receiving each treatment within each classroom were randomly selected for interviewing because it was not economically feasible to interview all students.

RESULTS

On each of the dependent variables at each grade level, a 3-way (treatment by sex by school) analysis of variance with unequal cell sizes was used to analyze the available data. Where the data were in the form of frequencies, Hartley's test (17) was run to test for homogeneity of variance. Where variances were significantly heterogeneous, log

TABLE 1

MEAN SCORES FOR MALES AND FEMALES AT EACH SCHOOL AND IN EACH TYPE OF ORIENTATION PROGRAM FOR EACH OF THE NINE CRITERIA

Criteria	Comprehensive Orientation Program											
	Male						Female					
	School*						School					
	1	2	3	4	5	6	1	2	3	4	5	6
Module test #1	14.3	15.0	16.0	11.6	40.4	33.9	15.2	16.3	15.0	13.6	36.6	38.0
Module test #2	8.3	6.9	7.7	7.8	**	7.2	8.9	7.2	6.9	8.4	**	6.8
Survey test	36.3	33.8	32.1	24.5	52.0	40.1	35.1	35.0	31.8	26.8	43.1	50.4
Opinion Survey	46.8	44.0	46.4	50.8	52.3	51.8	44.7	45.8	46.4	40.4	54.8	55.1
Structured Interview	15.0	17.0	14.8	12.0	12.9	16.0	15.0	14.6	14.0	10.0	10.5	13.9
# Math Modules	4.0	3.5	5.2	1.9	8.0	1.8	4.2	3.6	4.5	2.3	6.6	1.2
# Science Modules	1.4	1.3	3.5	2.6	3.4	2.5	1.7	1.6	2.8	2.7	2.0	1.8
# Soc Stud Modules	1.5	1.2	2.7	2.0	1.7	2.2	1.3	1.5	2.7	2.0	1.0	1.2
# L A Modules	3.9	2.9	6.2	3.5	2.8	2.2	5.3	4.2	6.3	3.4	2.5	1.6

Criteria	Brief Orientation Program											
	Male						Female					
	School						School					
	1	2	3	4	5	6	1	2	3	4	5	6
Module test #1	13.9	12.0	12.9	10.3	33.8	31.0	14.4	17.4	14.7	13.1	32.5	32.1
Module test #2	7.2	7.0	6.2	5.8	**	6.9	6.4	7.0	7.5	6.1	**	7.0
Survey test	33.7	29.5	28.7	24.9	46.1	43.0	37.4	33.4	33.3	27.1	49.3	44.2
Opinion Survey	44.8	35.5	42.6	41.6	54.6	52.9	47.3	46.4	47.3	40.8	58.1	51.2
Structured Interview	13.8	13.0	14.4	11.8	13.7	15.2	13.4	13.0	17.0	10.8	12.2	9.7
# Math Modules	3.3	3.5	4.3	2.8	7.7	1.5	4.0	2.9	5.1	2.1	6.5	1.1
# Science Modules	1.4	.5	3.1	2.9	3.8	2.8	1.7	1.3	3.1	2.8	2.5	1.6
# Soc Stud Modules	1.3	1.0	2.3	2.2	1.3	1.8	1.6	1.4	2.3	1.9	1.0	1.1
# L A Modules	3.3	3.0	6.9	3.6	2.9	1.8	4.2	2.7	6.7	3.6	2.6	1.

*   Schools 1-4 represent the four intermediate-level schools: schools 5-6 represent the two secondary-level schools.

** Blanks indicate no available data for that cell. Module test # 2 could not be administered at school 5.

TABLE 2

SIGNIFICANT F-SCORES*** AND THEIR CORRESPONDING DEGREES OF FREEDOM FOR THE MAIN AND INTERACTION EFFECTS FOR EACH OF THE NINE CRITERIA AT THE INTERMEDIATE LEVEL (IL) AND SECONDARY LEVEL (SL)

Sources of Variance	Criteria	Module test #1	Module test #2	Survey test	Opinion Survey	Structured Interview	# of Math Modules	# of Science Modules	# of Social Studies Modules	# of Language Arts Modules
Treatment (T)	IL		F=8.5* 1/102 df.							
	SL	F=22.9* 1/96 df.								
Sex (S)	IL	F=9.4# 1/165 df.			F=5.6** 1/154 df.					
	SL					F=8.2* 1/42 df.		F=12.5* 1/102 df.	F=15.8* 1/102 df.	
School (Sch)	IL	F=6.9* 3/165 df.		F=16.7* 3/155 df.	F=9.2* 3/154 df.	F=4.9* 3/70 df.	F=19.5* 3/165 df.	F=31.1* 3/165 df.	F=16.5* 3/165 df.	F=24.7* 3/165 df.
	SL	F=3.9** 1/96 df.					F=137.2* 1/102 df.	F=5.7* 1/102 df.		F=10.4* 1/102 df.
T x S	IL				F=6.6* 1/154 df.					
	SL									
T x Sch	IL									
	SL									
S x Sch	IL									
	SL	F=6.3* 1/96 df.		F=6.5* 1/87 df.						
T x S x Sch	IL			F=9.9* 1/87 df.						
	SL									

*p <.01
**p <.05
***F-scores that are underlined are for data that were transformed (log transformations)

transformations were made and Hartley's test was run again. In all but one case (i. e., the number of Science modules completed at the intermediate level) variances were, either before or after transformations, homogeneous.

In Table 1 are the mean scores obtained on the nine dependent variables by males and females from each school and in each type of orientation program. Table 2 shows the significant F-scores and their corresponding degrees of freedom for the main and interaction effects at each grade level for each of the nine criteria. Underlined F-scores indicate that the data were transformed by log transformations, as variances were found to be heterogeneous by Hartley's test.

The effect due to amount of orientation was significant in only two of eighteen F-tests. The students in the comprehensive orientation program performed significantly better on Module test number 1 (at the secondary level—SL) and Module test number 2 (at the intermediate level— IL) than did students in the brief orientation program.

With regard to the effects of Ss' sex, five of eighteen F-tests were significant. Female subjects performed significantly better on module Test No. 1 (IL) and had significantly more favorable opinions toward the PLAN program of individualized education (IL) than did male Ss. Male Ss performed significantly better during the interviews (SL) and completed significantly more Science modules (SL) and Social Studies modules (SL) than did females.

The effect due to school differences was significant in twelve of eighteen F-tests. There was significant variability among the schools on the following criteria (school rankings are based on the information in Table 1): Module Test No. 1(IL, School 2>3>1>4; SL, School 5>6) , Survey Test (IL, School 1>2>3>4); Opinion Survey (IL, School 3>1>4>2);Structured Interview (IL, School 3>2>1>4); number of Modules completed in Mathematics (IL, School 3>1>2>4: SL, School 5>6), in Science (IL, School 3>4>1>2: SL, School, 5>6), in Social Studies (IL, School 3>4>1>2), and in Language Arts (IL, School 3>1>2>4; SL, School 5>6).

Of the seventy-two effects only four were significant. The effect of the Ss' sex varied with the particular school they attended (i. e., on Module Test No. 1—SL and on the Survey Test—SL), with treatments on the Opinion Survey (IL) and with the second-order interaction of treatments and schools on the Survey Test (SL).

DISCUSSION

The results of this study provided little supporting evidence for the major hypothesis that students who participated in a comprehensive orientation program would perform better academically, have more knowledge of the educational system, and have more favorable opinions about the system than students who participated in a brief orientation program. Students in the comprehensive orientation program actually performed better than students in the brief orientation program on only the two Module Tests (No. 1 at the SL and No. 2 at the I L)

which reflected students' knowledge of the school system immediately after they completed their orientation activities. When comparing all students in the two programs, there were no differences on any other criteria.

The performance of male Ss differed from that of female Ss on several criteria. Other than attributing the results to possible maturation experiences, it is difficult to explain those findings which indicated that when males and females differed at the SL, males performed better; when they differed at the IL, females performed better. When the effects of schools were significant, students at one school consistently performed better than all other students, and students at a second school performed worse than all other students on most of the criterion measures. Specific differences in teacher abilities (e. g., the ability to conduct a PLAN classroom) might have contributed to these rather consistent school effects. Few interaction effects resulted and these were never significant at both the intermediate and the secondary academic levels.

The lack of evidence supporting the major hypothesis may have been related to several of the following factors. While there were extensive subjective data on which to base a belief that PLAN students had specific orientation needs, no objective measure of each student's needs was administered as a pretest device. If the orientation needs of PLAN students were different in kind or intensity from those anticipated, differential effects due to the amount or comprehensiveness of orientation perhaps should not have been expected.

It is also possible that the amount of orientation, as depicted by the comprehensive program was more than necessary. Perhaps the assumption that " the more information (or practice) the better " is not true with respect to orienting students. The best orientation program might be the one that provides each student with " just enough to get going in the system" and then allows the student to learn by performing in the system.

Since the students in the comprehensive and brief orientation programs were in the same classrooms, students in the two groups could have learned from one another, therefore confounding the effects of the treatments employed. Furthermore, the teachers may not have performed as the research design required. For example, they may have conducted group discussions in which material that was designated for the comprehensive program was covered with all the students. Or, students in the comprehensive program may not have participated in all the activities designated for their orientation program. Research assistants attempted to monitor teacher implementation of the research design but were unable to control and correct all deviations.

These results and implications suggest a number of guidelines for future research attempts to examine the effects of an orientation program for elementary and secondary students. First, the orientation needs of each student must be objectively assessed before, as well as after, the implementation of various programs or treatments. Secondly, in spite of the administrative difficulty involved, a no-orientation

control group should be employed to see if orientation of any kind is actually worthwhile. Thirdly, the effects of student role modeling and student interaction in the same classroom need to be either manipulated or specifically taken into account, perhaps by using measures of teacher characteristics so that all the students in a classroom could participate in the same program. Fourthly, criterion measures closely related to the orientation instructional objectives (and thus, to student orientation needs) should be employed. Hopefully, the objectives and the instruments will focus on more than student opinion and attitude variables. Finally, since it appears that the teacher makes a difference in the effectiveness of orientation activities, indices measuring factors which contribute to this difference should be incorporated into future research designs, (e.g., if how a teacher relates to and understands his students affects how his students learn, then this teacher quality or ability should be examined).

If subsequent research studies pursue these guidelines with more vigor than was possible in the present study, improved orientation programs should result. Only through these ways can questions such as the following be examined closely: (1) What student orientation needs exist for each student? (2) How can these needs best be met? (3) How can the degree to which these needs are met be determined?

This investigation was the first to examine orientation procedures in individualized educational systems and was an integral part of current developmental work on a comprehensive career guidance system (7) for such contexts. Also, it extended experimental research on orientation programs to academic levels below those of higher education, and it made progress toward using criterion measures directly related to the orientation instructional objectives and student needs. Subsequent research should extend these efforts and should incorporate the research design improvements suggested by the results of this study.

FOOTNOTES

1. The cooperation of Project PLAN teachers and administrators in the following California school districts is sincerely appreciated for the implementation of this study: Archdiocese of San Francisco, Fremont Union School District, San Jose Unified School District, and Santa Clara Unified School District. This study is part of a project being conducted pursuant to a contract with the Office of Education, United States Department of Health, Education, and Welfare (Contract No. OEG-0-070109-3530 (085), Research Project No. 7-0109).

2. G. Brian Jones, Daniel W. Kratochvil and Dennis E. Nelson are Director and Associate Research Scientists in the Guidance Research Program of the American Institutes for Research in Palo Alto, California. William E. Stilwell is Assistant Professor of Psychology at the University of Kentucky, Lexington, Kentucky.

REFERENCES

1. Cole C.; Ivey, A. E., " Differences Between Students Attending and Not Attending Pre-College Orientation," The Journal of College Student Personnel, 8:16-21, 1967.

2. Flanagan, J. C., Individualizing Education, American Institutes for Research, Palo Alto, California, 1968.

3. Foxley, C. H., "Orientation or Dis-Orientation?" The Personnel and Guidance Journal, 48: 218-221, 1969.

4. Gibbs, A., " Student Evaluation of Orientation," The Journal of College Student Personnel, 9: 158-160, 1968.

5. Hiehle, F. L., " Orientation to High School," Catholic School Journal, 68: 32-33, 1968.

6. Jessup, J. R., "Pre-College Orientation Conferences and Subsequent Behavior of Freshmen," The Journal of College Student Personnel, 7: 289-294, 1966.

7. Jones, G. B.; Nelson, D., Elements of a Comprehensive Guidance System Integrated in The Instructional Process, American Institutes for Research, Palo Alto, California, 1969.

8. Kiel, E. C., " College Orientation: A Disciplinary Approach," Liberal Education, 52:172-180, 1966.

9. McCann, C. J., " Trends in Orienting College Students," National Association of Women's Deans and Counselors Journal, 30: 85-90, 1967.

10. Miller, D. C.; Ivey, A. E., " Student Response To Three Types of Orientation Programs," Personnel and Guidance Journal, 45: 1025-29, 1967.

11. Nelson, E., " The Effectiveness of Freshman Orientation at Fourteen Colleges," School and Society, 55:138-139, 1942.

12. Nelson, E., " Measuring the Freshman Orientation Course," School and Society, 54:598-600, 1941.

13. O'Banion, T., " Experiment in Orientation of Junior College Students," The Journal of College Student Personnel, 10:12-15, 1969.

14. Packard, R. E., " Programmed Instruction Technique in New Student Orientation," The Journal of College Student Personnel, 9:246-252, 1968.

15. Pappas, J. G., " Student Reactions to a Small Group Orientation Approach," College and University, 43:84-89, 1967.

16. Rothman, L. K.; Leonard, D. G., " Effectiveness of Freshman Orientation," The Journal of College Student Personnel, 8:300-304, 1967.

17. Winer, B. J., Statistical Principles in Experimental Design, McGraw-Hill Book Co. 1962.

AGE, DEGREE OF TRAINING, AND TYPE

OF EXTRADIMENSIONAL SHIFT

IN NORMALLY INTELLIGENT HUMANS

MICHAEL D. LeBOW[2]
University of Manitoba

ABSTRACT

The experiment examined the relationship among several variables affecting extra-dimensional (ED) shift performance. Children and adults were trained to one of three criteria and given one of two ED shifts. All tasks required S to choose one of two colored geometric forms projected on a screen. The results are: (a) older Ss made fewer errors in training and transfer than younger Ss; (b) overtraining did not facilitate the ED shift wherein stimuli remained the same from training to transfer for the adults, but it did for the 7- and 8-year-old children; (c) overtraining inhibited the performance of the 5- and 6-year-old Ss given this ED shift; (d) overtraining neither facilitated nor inhibited the ED shift wherein stimuli changed from training to transfer. An interpretation was given in terms of verbal labeling, discrimination of change, perservative errors, and task difficulty.

FOR MANY YEARS psychologists studying the processes determining the course of human discrimination learning have used the transfer methodology. Conflicting results have been obtained especially with the extra-dimensional (ED) shift when training trials beyond criterion (overtraining) have been manipulated; this transfer task requires that the relevant training dimension become irrelevant during transfer. Caul and Ludvigson (1), for example, found that overtraining facilitated ED transfer learning with adults while Furth and Youniss (3), using normally intelligent children, showed that overtraining did not facilitate ED shift. Furthermore, Heal (6) found that overtraining inhibited ED transfer in a sample of retardates.

Several studies have raised the amount of original training beyond two values (i.e., criterion-training and overtraining) and have found that the relationship between degree of training and ED-shift errors is not linear. For example, Iwahara and Sugimura (7) administered an ED shift first to 4 and 5 year olds and then to 7 and 8 year olds while varying the amount of training between five and forty consecutive correct responses. Although the overall training effect was not significant, the authors found an inverted U shaped function with the ED-task becoming easier when the training criterion was increased beyond ten consecutive correct responses. With adults (15-19) a similar curvilinear relation has been found but with a maximum point lower than

ten (15). Speculating as to why these and other divergent results have been obtained, Wolff (17) pointed out differences between the ED-shift paradigms currently employed. The two most commonly used in overtraining experiments are (a) the ED-constant paradigm wherein cue values on the originally relevant dimension remain the same from training to transfer, and (b) the ED-change paradigm wherein cue values of the originally relevant training dimension are different in transfer. When overtraining has been found to facilitate ED-shift, the ED-constant procedure has generally been employed, but when no facilitation or inhibition has been found, the ED-change or some variant of it has been used. Unfortunately, the age of the Ss employed has been confounded with these paradigmatic differences. For example, when children up to the age of 8 are used as Ss, overtraining has not been found to facilitate ED-shift (3, 6). Eimas (2), however, did find that overtraining facilitated ED-shift in kindergarten and second grade children. According to Shepp and Turrisi's (12) interpretation of Zeaman and House's (19) theory, this contradictory result might be attributed to a strong irrelevant observing response facilitating ED-shift learning when the relevant observing response is weak. There is no evidence that these conditions existed in Eimas' (2) study. Furthermore, this feature of Zeaman and House's quantitative attention theory has never been tested (12). When adults are used as Ss overtraining has generally been found to facilitate ED-shift (1). While no

study has systematically attempted to manipulate age, degree of training, and type of ED-transfer at one time, several experiments have been reported in which some of these variables were separated. For example, there is suggestive evidence that overtraining does not result in large facilitating effects with ED-shifts involving cue changes along the formerly relevant dimension is Ss over 15 years of age (5). This is in contrast to the usually large effects obtained with adults given the ED-constant task. However, overtraining does not strongly facilitate the ED-constant task in children (7). Because the effect of intelligence on ED-transfer is also confounded with degree of training and type of shift (8), this variable was held constant in the study.

As an initial step in clarifying these issues, this experiment separated the effects of amount of training (5, 10, and 10+40) on ED-transfer (ED-constant and ED-change) in normally intelligent children (5-6, 7-8) and adults (18-21). Predictions were generated based on previous research and on the discriminable change hypothesis. This hypothesis considers that a diminution of shift task difficulty will occur as transfer becomes more discriminable (4, 13). According to this hypothesis, overtraining increasingly enhances the Ss discrimination of change of reinforcement contingencies involved in transfer. The readiness of any S to discriminate that a change in reinforcement contingencies has occurred may depend on the clarity of the training solution, and this clarity may increase with additional trials. Consequently, the retarding effect during ED-shift may be greatest with criterion trained Ss because they receive fewer trials. According to Wolff (17), the most substantial argument for the acceptance of this hypothesis is that it can predict the differential effect of overtraining on ED-constant and ED-change tasks, at least for adults.

H-1: ED-change (criterion 5) will be learned with fewer errors than ED-constant (criterion 5) and ED-change (criterion 10) will be learned with fewer errors than ED-constant (criterion 10) for all Ss. It is apparent that the ED-change shift is much more detectable than is the ED-constant shift. That is, changing the cue values of the previously relevant dimension (ED-change) makes the occurrence of transfer more obvious than in the ED-constant task, wherein no stimuli are altered.

H-2: Overtraining will facilitate ED-constant transfer in adults; i.e., ED-constant (criterion 10+40) will be easier than either ED-constant (criterion 5) or ED-constant (criterion 10).

H-3: ED-constant (criterion 5) will not be different from ED-constant (criterion 10) in adults. Because both of these criteria are sufficient for the adults to learn the task but not overlearn it, no difference between them is expected.

H-4: Overtraining will neither facilitate nor inhibit ED-change in any of the groups. The ED-change shift is assumed to make the change sufficiently discriminable that overtraining will not produce additional facilitative effects. Overtraining Ss should enhance the discrimination of change much more in situations wherein only reinforcement contingencies are changed (i.e., ED-constant), rather than in transfer tasks where both

cue values and reinforcement contingencies are altered (i.e., ED-change). That is, the change from training to transfer may be equally obvious to both criterion and overtrained Ss given the ED-change task. For the ED-constant, however, the change will be clearer for the overtrained.

H-5: Overtraining will not facilitate ED-constant in younger Ss (5 and 6, 7 and 8). While the effects of overtraining on ED-shift using children have not been systematically investigated, most of the studies have shown no facilitation. With the ED-constant task, in particular, overtraining seems to result in only weak facilitation. Perhaps, with young children, the additional trials employed have not been enough to enhance the discrimination of change between training and transfer.

METHOD

Subjects and Design

The Ss, seventy-two boys, were assigned in equal numbers to three different age groups. The oldest group of Ss ranged from 18 through 21 years of age (mean age 19 years 5 months). The mean IQ of this group was 103 excluding two persons who were dropped from the study because their IQ's markedly exceeded the normal limits. The next oldest group of Ss were between 7 and 9 years of age (mean age 8 years 2 months with a mean IQ of 96), and the youngest group were ages 5 and 6 (mean age 5 years 9 months with a mean IQ of 100). Two persons in the youngest age group were dropped from the study because they were unable to reach the training criterion before the eightieth trial. All Ss used in the study were of normal intelligence (i.e., IQ 90-110) as determined by scores of the Peabody Picture Vocabulary test administered at the time of the experiment. Most of the adult Ss who exceeded the age norms of the Peabody test were, in addition, given a Shipley Hartford Intelligence Test at the end of the study. Thus age (5 and 6, 7 and 8, and 18-21) constituted one independent variable in the design.

Amount of training prior to transfer was also manipulated. Equal number of Ss at each age level performed to a criterion (in terms of consecutive correct responses) of 5, 10, or 10+40 overtraining trials.

Type of ED-shift was the third variable in the design. The ED-constant shift involved the continued use of the stimuli (i.e., colored forms) employed in training while the relevant and irrelevant training dimensions were reversed. Under the ED-change shift, however, the cue values along the previously relevant training dimension were changed. Thus, for example, when square was positive and triangle negative during training, circle and cross replaced these cues during shift and color became the relevant dimension. To avoid any confounding of specific cues used in training with type of ED-shift, one half of the Ss were trained on a square versus triangle discrimination with circle versus cross replacing these cues in ED-change. The other half of the Ss were trained on a circle versus cross discrimination with square and triangle replacing these cues in ED-change. While this cue difference variable was factorially combined with all others, it was only a precaution against confounding and was not considered in any of the analyses.

The main experimental design was a 3x3x2 factorial having four Ss per cell with ages of Ss (5 and 6, 7 and 8, and 18-21), levels of training (5, 10, and 10+40), and type of ED-transfer (ED-change and ED-constant) serving as the independent variables. The major dependent variables used in the analyses were trails and errors to criterion after transfer.

Materials

The stimuli used in the experiment were geometric designs varying on two dimensions, each with two levels: form (square and triangle or circle and cross); color (black and red). These stimuli were prepared on 2-inch by 2-inch, 35 mm slides, with each slide being a photograph of two different forms of two different colors (e.g., black square and red triangle) and were randomly presented with the constraint that no identical slides appeared twice in succession. Two slide projectors, one for training and one for transfer, were used to project the stimuli on a translucent viewing screen located in front of the S. A 2-button response panel with transparent buttons was placed between the S and the translucent projection screen. A candy dispenser was located at the S's immediate right side, and a candy was dropped into a small paper container after every correct response. Stimulus presentation, switching from the first projector (training slides) to the second one (transfer slides), information feedback, and the dispensing of candy were controlled by a BRS solid state logic system. During the task, the E tabulated the data while sitting behind the experimental apparatus.

Procedure

Each S served individually and the instructions he received specified the details of the experiment as well as the response-reward contingency. All Ss were shown a sample training slide while the instructions were being read. In addition to the regular instructions, the 5 and 6 year olds were given a pre-training task to acquaint them further with the nature of the problem. Pilot data indicated that young children had difficulty understanding the regular instructions, especially the part pertaining to response and feedback. The pre-training stimuli consisted of several 35 mm slides, each being a photograph of two people or animals. The Ss learned to choose the stimulus which depicted the E's statement (e.g., "press the button underneath the mother" or "press the button underneath the man"). Five consecutive correct responses were sufficient to end pre-training and to begin the regular instructions, and after these were read, the task was begun.

Training. All the training tasks were form relevant, color irrelevant discriminations with either square or circle being the positive cue. For all Ss, red and black were the values of the irrelevant color dimension.

Transfer. Transfer automatically began as soon as the training criterion was reached. For all Ss, transfer consisted of a color relevant, form irrelevant discrimination with red being the positive cue and black the negative one. For half of the Ss, the previously relevant form training cues became irrelevant, but did not change in value (ED-constant). For the remaining half, these cues did change to new values on the form dimension (ED-change). The

transfer task continued until either ten consecutive correct responses were emitted or the eightieth trial was reached. At the end of the experiment, each S was allowed to keep his candies or cash them in for a penny each, whichever he preferred, and a brief interview concerning the S's ability to describe the experiment was administered.

During both training and transfer, the stimuli were projected on the translucent screen for an interval of time which was terminated by the S's response. Each S was required to select one of the two stimuli presented by pushing one of the two transparent buttons on the response panel. The S was instructed to push the button underneath the stimulus he thought was correct. Immediately after the Ss response, the stimuli were automatically removed from the screen and the button underneath the correct picture lit up for 5 seconds. Nine seconds following the S's response, a new pair of stimuli were presented. In addition, if the S was correct, candy was dispensed from the machine into the cardboard box. By answering the S's questions before the task began, conversation between the E and each S was minimized during the experiment.

RESULTS

The dependent variables of primary interest were trials and errors to criterion. Because these two measures were highly correlated (r = .95 and .97 for training and transfer, respectively), only the analyses of errors will be considered.

Training

The age by training by shift analysis of variance on number of errors to the fifth consecutive correct response is presented in Table 1. The analysis of variance revealed only an age effect, with the older Ss making fewer errors than the younger ones, (F [2, 54] = 3.8, p < .05, although the magnitude of this effect was not too large E^2 = .098). Age of the S was related to the ability to verbalize the correct stimulus-response (S-R) contingencies of training. From the interview administered at the end of the experiment, it was found that while twenty-four of the adults and seventeen of the 7 and 8 year olds could label the correct training cue, only two of the 5 and 6 year olds could accomplish this task. Errors were counted only to the fifth consecutive correct response to equate the different training levels. The probability of any S's making an error after the fifth consecutive

TABLE 1

SUMMARY OF ANALYSIS OF VARIANCE OF ERRORS IN TRAINING

Source	df	MS	F
Age	2	162.87	3.8153*
Training Level	2	114.67	2.6860
Shift	1	5.01	.1174
A x T	4	65.04	1.5235
A x S	2	42.35	.9919
T x S	2	7.06	.1652
A x T x S	4	27.64	.6474
Ss	54	42.69	
Total	71	46.97	

*p < .05

TABLE 2

SUMMARY OF VARIANCE OF ERRORS IN TRANSFER

Source	df	MS	F
Age	2	1377.06	11.6180**
Training Level	2	111.06	.9365
Shift	1	249.39	2.1040
Age x Training	4	375.37	3.1669*
Age x Shift	2	141.56	1.1942
Training x Shift	2	309.56	2.6116
Age x Training x Shift	4	215.03	1.8142
Ss/groups	54	118.53	
Total	71	181.55	

*p < .05
**p < .001

correct response was fairly low in all the conditions wherein the criterion of ten consecutive correct responses was required (i. e., only three of the 5 and 6 year olds, three of the 7 and 8 year olds and one of the 18-21 year olds made errors after achieving five consecutive correct responses).

Transfer

The age by training by transfer analysis of variance based upon errors made to the tenth consecutive correct response or the eightieth trial, whichever occurred first, is presented in Table 2. This analysis revealed an age effect (F [2, 54] = 11.6, p < .001) with younger Ss making more errors in transfer than older Ss. The magnitude of the relationship between age and errors in transfer was greater than in training (E^2 = .22). The youngest Ss also had difficulty naming the correct transfer cue at the end of the experiment, i. e., only three of the 5 and 6 year old group could state the appropriate transfer cue while fourteen of the 7 and 8 year old Ss and all of the adults could verbalize the correct solution. Furthermore, age and training interacted (F [4, 54] = 3.2, p < .05). The geometric representation of this interaction, presented in Figure 1, shows that while errors decreased for the 7 and 8 year olds from criterion 5 to criterion 10 + 40, they increased for the 5 and 6 year olds. Despite the lack of a significant F for an age by training by shift interaction, it might appear that there is a training by transfer interaction for the 7 and 8 year old groups. This F, however, was not significant (F [2, 18] = 2.3, p > .05). All the other main effects and interactions were not significant. In addition, it should be noted that while all the adults were able to reach criterion before the eightieth trial, several children could not (i. e., seven of the 5 and 6 year olds and four of the 7 and 8 year olds did not make ten consecutive correct responses in transfer).

Further Analyses of Errors in Transfer

Because hypotheses were made in advance of running the experiment, specific least significant difference (LSD) comparisons were made even when the overall F from the transfer analysis of variance was not significant (11). These LSD values are presented in Table 3.

TABLE 3

SUMMARY OF LEAST SIGNIFICANT DIFFERENCES FOR THE TRANSFER CONDITIONS

Hypothesis and Comparison	Difference	N	LSD
1. ED-change will be learned with fewer errors than ED-constant in the criterion trained groups			
ED-change (5)-ED-constant (5)	-134*	12	107.234
ED-change (10)-ED-constant (10)	+38	12	107.234
2. Overtraining will facilitate ED-constant in adults			
ED-constant (10+40)-ED-constant (10) 18-21 years only	-4	4	61.707
ED-constant (10+40)-ED-constant (5) 18-21 years only	+8	4	61.707
3. The criterion ED-constant conditions will not be significantly different in adults			
ED-constant (10)-ED-constant (5) 18-21 years only	+12	4	61.707
4. Overtraining will not facilitate ED-change			
ED-change (10+40)-ED-change (10)	-33	12	107.234
ED-change (10+40)-ED-change (5)	+95	12	107.234
5. Overtraining will not facilitate ED-constant in children			
ED-constant (10+40)-ED-constant (10) 7 and 8 years only	-29	4	61.707
ED-constant (10+40)-ED-constant (5) 7 and 8 years only	-83*	4	61.707
ED-constant (10+40)-ED-constant (10) 5 and 6 years only	+76*	4	61.707
ED-constant (10+40)-ED-constant (5) 5 and 6 years only	+74*	4	61.707

*p < .05

ED-constant, ED-change, and Criterion Training. It was found, as hypothesized, that for the criterion

5 conditions, fewer errors were made in the ED-change than ED-constant task for all Ss ($p < .05$). This was not the finding, however, for the criterion 10 conditions. While the data showed that more errors were committed in ED-change (10) than in ED-constant (10), this difference was not reliable.

ED-constant, ED-change, and Overtraining. Overtraining did not facilitate the ED-constant task in adults. That is, when the ED-constant (10+40) condition was compared to both the ED-constant (10) and the ED-constant (5) conditions, for the adult group, no significant differences were found. Furthermore, ED-constant (10) was not significantly different from ED-constant (5). In short, it appears that the amount of original training had no appreciable effect on the number of errors made in the ED-constant task for the adult group. As hypothesized, overtraining did not significantly facilitate or inhibit ED-change performance for all the different age groups combined.

ED-constant, ED-change and Overtraining in Children. It was hypothesized that overtraining would not facilitate the ED-constant task in children. It was found, however, that 7 and 8 year olds in the ED-constant overtraining condition made fewer errors than in both of the other ED-constant tasks with the difference between the ED-constant (5) and ED-constant (10+40) conditions proving reliable ($p < .05$). The opposite was found for the 5 and 6 year olds, with more errors being made in the ED-constant overtraining transfer condition than the other two

FIGURE 1.

AGE BY TRAINING INTERACTION

AGE X TRAINING

transfer groups combined ($p < .05$ for both). Interestingly, when both the ED-constant and ED-change transfer conditions were combined for the 5 and 6 year olds, a positive linear trend was found in the number of errors made from the lowest to the highest training level. The nonlinear portion of this trend was not significant.

DISCUSSION

While it was found that ED-change was learned with fewer errors than ED-constant shift for the criterion 5 conditions and that overtraining neither facilitated nor inhibited ED-change learning, the remaining major predictions made for this study were not confirmed.

Degree of Training and ED-Change

It was predicted that changing the cues of the previously relevant training dimension in transfer would facilitate learning in the criterion trained groups in comparison to no change at all. Furthermore, it was predicted that this facilitation from changing the cue would mask any facilitation from overtraining; that is, the switch from training to transfer would be equally obvious for both the criterion and overtrained Ss given the ED-change task. It was found, however, that only with the lowest training level (criterion 5) was ED-change easier than ED-constant. With the criterion 10 conditions, ED-constant was easier than ED-change although this difference was not reliable. The difference between the criterion 10 conditions is mostly within the 5 and 6 year old age group. The 7 and 8 year old Ss performed in the predicted direction showing fewer errors in ED-change than in ED-constant criterion 10. It is not clear why 5 and 6 year olds emitted more errors in the ED-change task than in the ED-constant task under the criterion 10 condition.

The effects of overtraining on ED-change were consistent with the prediction. Other experiments have shown that a high degree of original training either does not facilitate or mildly interferes with the ED-change task in children (18). This was found for both groups of children in the present study when the ED-change overtraining condition was compared to the lowest training criterion (5); however, only the 7 and 8 year old Ss found ED-change overtraining more difficult than the ED-change criterion 10 condition. It is interesting that the performance of these Ss in ED-change and ED-constant is noticeably different. Increasing the training trials for 7 and 8 year olds seems to facilitate ED-constant but inhibit ED-change performance. While the performance of 5 and 6 year old Ss under ED-change was different from their performance under ED-constant, in both of these transfer conditions overtraining was harder than the lowest training level.

For the adult group the overtraining ED-change task was the easiest discrimination, although not much different from the criterion 5 condition. Both the form and magnitude of the curves for ED-change and ED-constant, which depict errors in transfer for the different training levels, were similar, indicating that, for these Ss, cue change was not a crucial variable.

While no definite statement can be offered
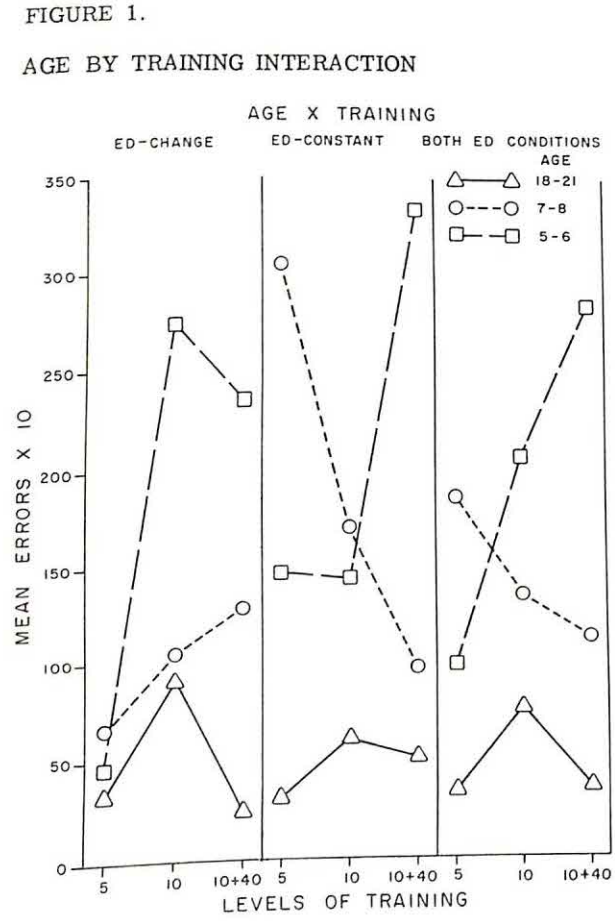
explaining these results with children, it does seem that changing the cues along the previously relevant training dimension in transfer affects the performance of these Ss.

Degree of Training and ED-Constant Transfer in Adults

Contrary to the third prediction, overtraining did not facilitate ED-constant transfer in adults. Evidently, a large amount of original training did not significantly affect these Ss in ED-constant shift. Perhaps the ED-constant discrimination was sufficiently easy for adult Ss, so that the facilitating effects of overtraining would not be readily apparent. Low levels of training, in this case, would be enough to ensure rapid solution in transfer. In short, each of the training levels was stringent enough to alert Ss that errors, in the initial phase of the shift, were indicative of a change in procedure rather than mistakes of the E or themselves. At the very least, it can be stated that large amounts of training did not increase the ease of discrimination of a change in procedure for these Ss. In addition, all of the adult Ss were able to verbalize the correct stimuli for training and transfer. Therefore, it seems reasonable to conclude that the adults trained to criteria of five and ten consecutive correct responses were highly trained, and adding the extra training trials did not enhance the learning of the ED-constant task.

The ease of the training and transfer discriminations may be related to the S's ability to label the stimulus cues and S-R contingencies. Kendler and Kendler (9) suggest that training in a discrimination situation establishes mediating responses to the dimensional aspect of the physical stimulus. The addition of response-produced cues facilitates instrumental response learning. For humans, these implicit response-produced cues may be verbal labels. In a simple 2-dimensional discrimination, adults can label the stimulus and S-R contingencies after only a few correct trials. Perhaps, well developed labeling abilities enhance the discrimination of change in reinforcement contingencies. This discrimination of change may function to reduce perseverative errors to previously relevant training cues and/or the previously relevant training dimension leading to a more rapid transfer solution. In other words, by more quickly recognizing incorrect responses as indicants of a change in task rather than haphazard mistakes of the E or themselves, Ss will learn the ED-constant task with fewer errors. Since the present experiment did not allow an adequate test of perseverative responses in transfer, the hypothesized relationship between discrimination of change and perseverative errors is not derivable from the results of this study.

Degree of Training and ED-Constant Transfer in Children

For the 7 and 8 year old Ss, overtraining was found to have a facilitating effect on ED-constant learning. As the degree of original training increased, the errors made in ED-constant transfer decreased with the ED-constant overtraining condition being easier than either of the criterion trained groups.

By employing the discriminable change hypothesis,

the present results found for the 7 and 8 year old Ss in the ED-constant task can be explained. Simply stated, the effects of large amounts of original training might have been to enhance the clarity of the switch in reinforcement contingencies leading to a more rapid solution of the ED-constant task. If the training and ED-constant discriminations were at an optimal level of difficulty for these Ss, such that they could attach labels to the stimuli and the correct S-R contingencies, then additional training trials (by enhancing the discrimination of change between the two tasks) would have a facilitative effect. Indeed, as the training criteria were raised, these Ss made progressively fewer errors in ED-constant transfer.

Another somewhat related explanation is the 2-factor theory of discrimination learning (7, 8, 14, 15, 16). According to this approach, discrimination learning involves S-R connections in the early stages and the acquisition of general discrimination sets in the later stages. Discrimination sets, in contrast to learning sets, can be acquired in one task. The specific or single unit S-R connections lead to negative effects in ED-transfer, while the discrimination sets have positive effects.

According to Iwahara and Sugimura (8), the acquisition of discrimination sets is correlated with intelligence of the S and the degree of original training. With low training criteria, the discrimination set may not be acquired. The fact that a shift may become more difficult with an increase in training up to a certain point is accounted for by the negative transfer effects of the specific S-R connections acquired in training. Conversely, the decreasing tendency in errors during shift with an increase in training trials beyond that certain point is due to the facilitative effects of discrimination sets also acquired in training. While these authors have not explained how a discrimination set facilitates ED-constant learning, or for that matter what constitutes a discrimination set, it may well be a labeling process. That is, as the training trials are increased, Ss may become progressively better at naming the stimuli and correct S-R contingencies resulting in a better discrimination of change when transfer occurs. Thus, a discrimination set may be what Kendler and Kendler (9) call verbal labels. According to these authors, verbal labeling abilities should be fairly well developed in humans above the age of 6. To repeat, effective labeling may function to enhance the discrimination of change from training to transfer. This labeling process is facilitated by additional training trials and may lead to a reduction in perseverative errors of both kinds (e.g., previously relevant cue and dimension). However, if the discrimination employed becomes too difficult and Ss are unable to effectively label, or the Ss are deficient in labeling abilities, then additional training trials may have adverse effects.

In contrast to the results found for the older children, 5 and 6 year old Ss made significantly more errors in ED-constant transfer when overtrained than when required to emit either five or ten consecutive correct responses during training. The differentive difference between the two criterion trained ED-constant conditions was slight. The fact that overtraining inhibited ED-constant transfer in 5 and 6 year olds and facilitated it for 7 and 8 year olds may also be related to the ability to label. From the interviews

given at the end of the experiment, it was found that only about 10 percent of the 5- and 6-year-old Ss were able to name the correct training and transfer stimuli, while almost 65 percent of the 7 and 8 year old group could perform this task. Kendler and Kendler (10) have suggested that children below the age of 6 are deficient in the ability to form verbal mediating responses. If being able to name the stimuli and contingencies was difficult for these Ss, then overtraining would not be expected to facilitate ED-constant learning. Furthermore, if, in fact, these Ss learned the problem in a rote manner (e. g., based upon single-unit S-R principles) then overtraining might, as suggested, retard ED-constant learning. Iwahara and Sugimura (8) found that overtraining feebleminded adolescents resulted in a retardation of ED-constant shift learning as compared to low levels of training. The ED-constant shift was facilitated after overtraining for the more intelligent Ss. The authors concluded that the feebleminded Ss learned the problem more mechanistically than conceptually; partial reinforcement of the previously relevant training cue in ED-constant transfer might enhance perseverative errors to a greater extent, if the S learned the training problem in a rote manner. As suggested previously, learning a discrimination problem mechanistically may involve the absence of, or a reduction in, the effectiveness of being able to label the stimuli and S-R contingencies. This would function to reduce the discrimination of change and enhance the negative effects in ED-constant transfer found after large amounts of training. That is, with poor discrimination of change resulting from impoverished labeling abilities perhaps many perseverative errors would be committed in ED-constant transfer. In this situation, possibly because of the greater number of extinction trials needed and the exacerbating effects of partial reinforcement of previously relevant training responses, more errors would be made after overtraining than criterion training. The feebleminded Ss of Iwahara and Sugimura's (8) study as well as the 5 and 6 year old Ss in the present experiment may have both been unable to effectively label the stimuli and thus performed more poorly in ED-constant after overtaining. In sum, it has been suggested that a large amount of original training would facilitate ED-constant learning if Ss had fairly well developed labeling abilities and inhibit this transfer task if Ss were deficient in this ability. Furthermore, if the training and transfer tasks were easy enough for sufficient labeling to take place after low levels of training, then the facilitating effects of overtraining in ED-constant transfer would be negligible.

CONCLUSIONS

While most of the main predictions of this study were not confirmed, several conclusions are warrented: (a) age of the S appears to be a crucial variable affecting shift; (b) for adults, degree of training will not be a crucial variable in ED-constant learning if the transfer discriminations are too simple; (c) degree of training may have different effects in older, in contrast to younger, children given the ED-constant task. It was suggested that these differences may be related to the ability to label the stimuli and S-R contingencies; and (d) overtraining did not significantly facilitate or inhibit ED-change performance. However, it seems that under at least some training levels, ED-change and ED-constant have different effects as a function of age.

An investigation of the effects of labeling stimuli and S-R contingencies in training and ED-constant transfer after different amounts of training would be useful. In this context, it would also be important to investigate the relationship of labeling, discrimination of change between training and transfer, and the commission of perseverative errors to the previously relevant training cue and dimension.

It seems reasonable that labeling and utilizing the necessary information in a discrimination problem is, in part, related to the difficulty of the problem (e. g., number of dimensions, kinds of stimuli used, and type of shift). It also appears reasonable that partial training for 5 and 6 year old Ss may be sufficient for 7 and 8 year olds to learn the task, and may be overtraining for adults, if the same discrimination were employed. Equating for task complexity across the different age groups might result in similar performance under each training level for easy tasks, where labeling is not difficult and gross differences appear with the more complex discriminations. However, to delineate the relationship between task complexity and the ability to label, as it influences the discrimination of change between training and ED-constant transfer, requires further research.

REFERENCES

1. Caul, W. F.; Ludvigson, H. W., "The Effect of Overlearning on Response Shifting, *Psychonomic Science*, 1:159-160, 1964.

2. Eimas, P. D., "Effects of Overtraining and Age on Intradimensional and Extradimensional Shifts in Children," *Journal of Experimental Child Psychology*, 3:348-355, 1966.

3. Furth, H. G.; Youniss, J., "Effect of Overtraining on Three Discrimination Shifts in Children," *Journal of Comparative and Physiological Psychology*, 57:290-293, 1964.

4. Grant, D. A.; Cost, J. R., "Continuities and Discontinuities in Conceptual Behavior in a Card Sorting Problem," *Journal of General Psychology*, 50:237-244, 1964.

5. Guy, D. E.; Van Fleet, F. M.; Bourne, L. E., Jr., "Effects of Adding a Stimulus Dimension Prior to a Nonreversal Shift," *Journal of Experimental Psychology*, 72:161-168, 1966.

6. Heal, L. W., "The Role of Cue Value, Cue Novelty, and Overtraining in the Discrimination Shift

Performance of Retardates and Normal Children of Comparable Discrimination Ability," Journal of Experimental Child Psychology, 4:125-142, 1966.

7. Iwahara, S.; Sugimura, T., "Studies on Shifts of Discrimination Learning: I. The Number of Trials During Prior Learning," Japanese Journal of Educational Psychology, 6:106-112, 135-136, 1958.

8. Iwahara, S.; Sugimura, T., "The Effect of Different Amounts of Training Upon Shifted Discrimination Learning as a Function of Intellectual Level," Psychologia, 5:93-98, 1962.

9. Kendler, H. H.; Kendler, Trace S., "Vertical and Horizontal Processes in Problem Solving," Psychological Review, 69:1-16, 1962.

10. Kendler, Trace S.; Kendler, H. H., "Reversal and Nonreversal Shifts in Kindergarten Children," Journal of Experimental Psychology, 58:56-60, 1959.

11. Li, J. C. R., Statistical Inference: I, Edwards Brothers, Inc., Michigan, 1964.

12. Shepp, B. E.; Turrisi, F. D., "Learning and Transfer of Mediating Responses in Discrimination Learning," Ellis, N. R.(ed.), International Review of Research in Mental Retardation, Academic Press, New York, Volume 2, 1967.

13. Slamecka, N. J., "A Methodological Analysis of Shift Paradigms in Human Discrimination Learning," Psychological Bulletin, 69:423-438, 1968.

14. Sugimura, T., "Effect of Degree of Original Learning and Time Interval Between Two Tasks on Response Shift, Japanese Journal of Psychology, 33:133-140, 1962.

15. Sugimura, T.; Iwahara, S., "Studies on Shifts of Discrimination Learning: II Effect of a Wide Range of Trials in Prior Learning," Japanese Journal of Educational Psychology, 7:142-147, 191, 1959.

16. Suzuki, S., "Study on Shifts of Discrimination Learning in Children," Japanese Journal of Educational Psychology, 9:84-91, 127-128, 1961.

17. Wolff, J. L., "Concept-shift and Discrimination Reversal Learning in Humans," Psychological Bulletin, 68:369-408, 1967.

18. Youniss, J.; Furth, H. G., "Discrimination Shifts as a Function of Degree of Training in Children," Journal of Experimental Psychology, 70:424-427, 1965.

19. Zeaman, D.; House, Betty J., "The Role of Attention in Retardate Discrimination Learning," Ellis, N. R. (ed.), Handbook of Mental Deficiency, Psychological Theory and Research, McGraw-Hill, New York, 1963, pp. 159-223.

EMPIRICAL EVIDENCE ON THE

APPLICATION OF LORD'S SAMPLING TECHNIQUE

TO LIKERT ITEMS[1]

RICHARD C. PUGH
Indiana University

ABSTRACT

This study was conducted to evaluate empirically if means and standard deviations could be estimated accurately by Lord's technique for an attitude scale which consisted of Likert items. The procedures followed were those described by Lord (2). The formulas used were those reported by Plumlee (3), except that a minor modification was made in the formula for estimating the population standard deviation. An estimate of the population mean using item sampling procedures based on 5 percent item samples (3 items in each sample) was less deviate from the population mean than fifteen of the twenty examinee sample estimates (5% examinee sample). A 10 percent item sample estimate of the population mean was less in error than nine of the ten corresponding examinee sample estimates and a 20 percent item sample estimate was more accurate than all five of the estimates from the 20 percent examinee samples. A 5 percent item sample estimate of the population standard deviation was more accurate than twelve of the twenty examinee sample estimates. A 10 percent item sample estimate was in less error than all ten of the examinee sample estimates and an estimate from the 20 percent item sample was more accurate than four of the five examinee sample estimates. Usable estimates of the population mean and standard deviation seemed to be obtained.

LORD'S (2) technique for the development of test norms by using item sampling procedures has been empirically checked for tests that are scored by number right. Lord (2), Plumlee (3), and Cook and Stufflebeam (1) as well as others have shown evidence that item sampling is as effective as examinee sampling, if not more so, in test norming. The advantage in item sampling for each student to spend only a few minutes to answer a few items instead of many minutes to answer an entire test has appeal in school districts where new instruments are being used or developed.

This study was conducted to evaluate empirically if means and standard deviations could be estimated accurately by Lord's technique using an attitude scale which consisted of Likert items. Each item was scored on a 4-point continuum: 4 points for strongly agree, 3 points for agree, 2 points for disagree, and 1 point for strongly disagree. The need for the study came from the use of an attitude scale in a school district to evaluate an exemplary program. The item sampling technique was more feasible than traditional approaches in estimating the population mean and standard deviation but no previous studies could be found except ones in which the tests were scored by number right.

METHODS AND DATA SOURCES

The procedures followed were those described by Lord (2) and the formulas used were those reported by Plumlee (3) except that a modification was made in the formula for estimating the population standard deviation in order to adapt it to Likert items. An available Likert type attitude scale in the area of student satisfaction with school was utilized. This scale consisted of sixty items and was given to a population of six hundred fifth and sixth grade students. The scale was not timed and approximately 1 hour was allotted for students to respond to the items on the scale.

Item sample and examinee sample estimates of the population mean and standard deviation were made from item and examinee samples of 5, 10, and 20 percent. The examinee samples were drawn at random without replacement and consisted of twenty random samples (30 examinees each) for the 5 percent samples, ten random samples (60 examinees each) for the 10 percent samples, and five random samples (120 examinees each) for the 20 percent samples. In a similar manner, the item samples were formed by random sampling without replacement and each item sample consisted of three items for the 5 percent samples, six items for the 10

percent samples, and twelve items for the 20 percent samples.

Means, standard deviations, and item variances were computed for each examinee and item sample. The means, standard deviations, and item variances for the item samples were used to compute an estimate of the population mean and an estimate of the population standard deviation. The formulas presented by Plumlee (3) were used except that the item variance sum term (Σpq) of the standard deviation formula was converted to its algebraic equivalent for Likert items scored on a scale from 1 to 4, Σx_i^2, in which m stands for the number of examinees $\overline{m-1}$ taking a subtest.

These data were obtained from students attending elementary school adjacent to the Bloomington campus of Indiana University, a community consisting of an above average number of professionally oriented families.

RESULTS AND CONCLUSIONS

The estimate of the population mean using the item sampling procedure based on the 5 percent item samples (3 items in each sample) was less deviate from the population mean than fifteen of the twenty examinee sample estimates (30 examinees in each sample). The 10 percent item sample estimate of the population mean was less in error than nine of the ten examinee sample estimates. The 20 percent item sample estimate of the population mean was more accurate than all five of the estimates from the 20 percent examinee samples.

The relative accuracy of the estimate of the population standard deviation from the item samples was similar to the estimates of the population mean when examinee sample estimates were used for comparison. For the 5 percent sample estimate, the item sample estimate of the population standard deviation was more accurate than thirteen of the twenty examinee sample estimates. The 10 percent sample estimate was in less error than all ten of the examinee sample estimates, and the estimate from the 20 percent item sample was more accurate than four of the five examinee sample estimates.

EDUCATIONAL IMPORTANCE OF THE STUDY

For the data obtained from this 60-item Likert type attitude scale, the estimation of the mean by item sampling was close to the population mean, not deviating more than 1.3 raw score points from a population mean of 175.45 (see Table 1). Compared to examinee sample estimates, the item sample estimates seemed to have an advantage. The item sample estimate of the population standard deviation of 25.47 was in error 1.53 raw score points for the 5 percent item sample estimate, 0.08 points for the 10 percent estimate, and 0.55 points for the 20 percent estimate. Based on these data, the standard deviation was fairly well estimated; however, no information was obtained on the shape of the distributions and estimates of norm tables were not studied.

If it is more feasible in some schools for each student to spend only a few minutes responding to a

TABLE 1

COMPARISON OF OBTAINED POPULATION MEAN AND STANDARD DEVIATION WITH THOSE ESTIMATED BY ITEM-SAMPLING AND EXAMINEE-SAMPLING METHODS ON A 60-ITEM LIKERT TYPE ATTITUDE SCALE

Method and Sample	N	Mean	Standard Deviation
Population	600	175.45	25.47
5 Percent Examinee Sample Estimates			
A	30	169.77	31.21
B	30	178.60	22.62
C	30	180.30	22.40
D	30	172.83	30.79
E	30	176.27	23.76
F	30	170.80	28.39
G	30	174.50	22.94
H	30	171.80	27.78
I	30	176.80	25.25
J	30	174.70	27.50
K	30	177.10	24.83
L	30	176.50	19.14
M	30	179.30	25.53
N	30	176.50	24.15
O	30	178.93	22.70
P	30	178.63	20.25
Q	30	168.03	24.77
R	30	169.37	27.74
S	30	179.10	23.97
T	30	179.17	24.64
5 Percent Item Sample Estimate (3 items each)			
A-T	30/sample	174.17	23.94
10 Percent Examinee Sample Estimates			
A	60	176.08	25.38
B	60	172.95	28.29
C	60	177.85	21.70
D	60	180.17	24.12
E	60	173.90	26.77
F	60	178.97	24.34
G	60	173.58	24.20
H	60	170.93	23.80
I	60	171.65	28.47
J	60	178.42	24.95
10 Percent Item Sample Estimate (6 items each)			
A-J	60/sample	174.42	25.39
20 Percent Examinee Sample Estimates			
A	120	177.53	24.91
B	120	173.27	26.33
C	120	174.39	23.04
D	120	175.91	26.73
E	120	176.16	25.98
20 Percent Item Sample Estimate (12 items each)			
A-E	120/sample	175.48	26.02

few items instead of a much longer time for an entire attitude scale, then item sampling procedures offer some help as usable estimates of the population mean and standard deviation seem to be obtained.   However, further verification of the procedure for Likert type attitude scales needs to be made as this study was completed on a single, 60-item scale in which each item was scored on a range from 1 to 4 points.

FOOTNOTES

1. An earlier version of this paper was presented at the 1970 Annual Meeting of the National Council on Measurement in Education, Minneapolis, Minnesota, March 1970.

REFERENCES

1. Cook, Desmond L.; Stufflebeam, Daniel L., "Estimating Test Norms From Variable Size Item and Examinee Samples," Educational and Psychological Measurement, 27:601-610, 1967.

2. Lord, Frederic M., "Estimating Norms by Item-Sampling," Educational and Psychological Measurement, 22:259-267, 1962.

3. Plumlee, Lynnette B., "Estimating Means and Standard Deviations from Partial Data—An Empirical Check on Lord's Item Sampling Technique," Educational and Psychological Measurement, 24:623-630, 1964.

BOOK REVIEWS

Robert E. Clasen

book review editor

NEW DIMENSIONS IN HIGHER EDUCATION

Collier, K.G. (New York: Humanities Press, Inc., 1968), 164 pp.

THE AUTHOR has had most of his teaching experience in departments and schools of education in apprentice colleges. He has taught courses in education for a 6-week term of summer school at New York University. He is now and has been for a number of years an instructor and professor at the Venerable Bede College of the University of Durham and has, in recent years, been principal of that college. Naturally, the book is oriented toward higher education in Great Britain, but the author is apparently very well informed about higher education in the United States.

There are two new dimensions discussed in the book. The first is the greater scale on which full-time higher education is likely to be conducted within the next generation. By the end of the century, perhaps one in three, or even one in two, in the immediate post-high-school age group will be participating in higher education. This compares with one in twelve in 1962. The second new dimension is that opened up by educational technology, with the sharper awareness it is forcing on us of what higher education is and what it might be.

The discussions are obviously for liberal arts colleges in higher education rather than other types of higher education, and the author commits himself to the fundamental premise that a liberal arts education should be general education rather than specialized education. In Collier's opinion, the type of subjects which may well be offered for general education should be governed by five principles:

1. Contemporary civilization is inescapably dependent on science and technology and the modern citizen needs to appreciate the positive use both of scientific methods and of technological development in the normal life of the country.

2. Modern society is changing more rapidly than any previous human society. World population is increasing at phenomenal speed; communities of all kinds, whether urban or rural, industrial or underdeveloped, are growing in scale and changing in style in unforeseeable ways.

3. We live in a society where there is no longer any clear consensus among leading citizens as to the values that should rule our lives. The older generation find it difficult to help young adults to form their values. The speed of social change makes adaptiveness and flexibility more important than in the past.

4. There must be some continuing study of language and its assumptions, and practice in its use.

5. There must be provision for synthesis, since students will be following not only a specialist course, but non-specialist courses on "science and society," "social change," and "values." This is perhaps best done through syndicate (small group) assignments and seminars conducted by tutors who are familiar with the general courses and are concerned to attempt some synthesis.

Throughout the book, the author discusses various types of teaching methods which satisfy four basic conditions for good teaching: (1) Intense activity of the mind of the learner in a continual process of comparing and

PROGRAMMED TUTORING OF DECODING SKILLS

WITH THIRD AND FIFTH GRADE NON-READERS

ELLIS RICHARDSON and LUCY COLLIER
New York State Department of Mental Hygiene, Child Psychiatric Evaluation Unit

ABSTRACT

The acquisition of decoding skills (sound-symbol correspondence, visual analysis, and blending) was studied with twelve Ss who scored below average on a battery of psychomotor tests. A group of twelve " no treatment " control Ss were shown to be superior to the experimental Ss in reading simple sight words on a laboratory pretest. Each experimental S required an average of 4 1/2 hours of tutorial time, distributed across forty-three sessions, in learning the program content. Posttest results showed the experimental Ss to be superior to the controls on all measures of decoding and demonstrated that experimental Ss could apply decoding skills to unfamiliar content. The major conclusion drawn is that so-called dyslexic children can learn basic reading skills. The success was attributed to the highly-structured, programmed approach.

ALTHOUGH MUCH has been published regarding the nature and cause of reading disabilities (13), little is known about effective approaches to the correction of these problems. Studies implicate emotional maladjustment (11:74), cultural asynchrony (18), and neurological impairment (13:57-71). Remedial approaches range from physical therapy (4), multi-sensory stimulation (Vakt and Fernalt tracing techniques) (15:43), and specific sensory modality training in deficit perceptual areas (15:43-44), to varied techniques of teaching reading itself.

Progress has been made in identifying the parameters which will predict reading failure at the readiness stage (5), yet no effective program has been documented which provides the remedial help for children who need it early in school. In 1968, one-third of New York City's public school children between the second and ninth grades were reading one year behind the national norm, and one-fourth were reading 2 years behind the national norm, as measured by the Metropolitan Reading Achievement Test (MRAT) (12). But the standardized scores do not accurately reflect the number of children severely retarded in reading (16), and it is frequently for those children least responsive to remedial intervention that no help is available within the educational institution.

It is clear from a recent survey of reading studies (2:103, 107) that a " code emphasis " method, one that emphasizes consistent training in decoding printed language for spoken language, in contrast to a " meaning emphasis " method, produces better results at the beginning reading stage.

The decoding skills outlined by recent investigators (7, 10, 14) provide the behaviors which allow a child to determine the word-sound corresponding to a word-image through a knowledge of its sound components. The usual look-say skill requires that a child learn a separate sound association for each word configuration. The decoding approach requires that the reader first analyze a word into its parts on a visual basis (e. g. , for the trigram man, one of three divisions are possible: m, a, n; m, an; ma, n) sounding out the word components in left-to-right order with the knowledge of single letter and bigram sound associations. Then, he must be able to derive the oral blend for the whole word from the component parts. The analytical reader then has a tool for decoding new letter combinations made up of familiar parts. Research has indicated that the average child cannot induce letter-sound and blend-sound relationships from look-say training without explicit training in decoding skills (14:30).

Chall, Roswell, and Blumenthal (3) have shown that auditory blending ability is positively correlated with reading achievement on several different

measures (silent reading, oral reading, and phonic ability). Silberman (14) has explored this relationship. He investigated techniques for teaching analytical decoding skills with consonant-vowel-consonant (CVC) trigrams and found that the key to training the generalized skill for reading novel trigrams was auditory blending practice. First grade children who were not trained to respond with a whole word-sound (e. g. , man) to the auditory presentation of the phoneticized elements (e. g. , /m/, /an/) were not able to apply their analytical skills to novel combinations of trigram elements. His Ss who had been trained in the auditory blend were able to decode 75 percent of the novel trigrams tested, while Ss who did not receive auditory training in an earlier version of the program were unable to read any of the novel trigrams.

Gotkin, McSweeney, and Richardson (7) have recently developed a similar program for kindergarten children. They found that kindergarten children could generalize to one-third of the seven novel trigrams tested after completion of the program. They found it necessary to place even heavier emphasis on the auditory blending training than did Silberman (14). The Gotkin study generated evidence that further training in the programmed routines would increase the percentage of novel words a child could read.

Silberman and Gotkin demonstrated that decoding skills could be taught to first grade and kindergarten populations with automated individualized lessons. The project described here represents an attempt to study in detail the acquisition of essential decoding skills by third and fifth grade nonreaders who were not benefiting adequately from remedial help available in the school. The program extended and revised the Gotkin work and utilized a programmed tutorial technique similar in many aspects to techniques developed by Ellson and others (6).

DESCRIPTION OF THE PROGRAM

The most important behavioral objectives of the program may be summarized in the following model (10:28):

1. Child looks at " pom."
2. Child perceives " p" and " om " as separate units.
3. Child says sounds in order, " /p/, /om/. "
4. Child listens to the sounds he has verbalized.
5. Child produces sound for the whole image, " pom. "

In the first three steps, the child must have the sound-symbol associations for the images " p" and " om ", he must be able to perceive them as ordered units of the whole image, " pom," and he must be able to produce their sounds in order. These steps are based on the visual modality. In steps 4 and 5, the child must be able to blend his own verbalization of two discrete sounds into a single composite sound. These steps are based on the auditory modality. The objectives, then, define a word-attack behavior which, in its generalized form, can be used by the child to unlock the word-sound of any regular bigram or trigram. Other behavioral objectives include a small sight-word vocabulary, capital letters, labeling punctuation marks, and proper inflection in oral reading.

This program is organized into three cycles or sets of lessons with seventeen lessons in Cycle I and ten lessons each in Cycles II and III. Each lesson is preceded by about 1 minute of sound training in which the child practices saying the new sounds or words to be introduced in the lesson. For lessons specially designed to teach the blending skill, each lesson is preceded with practice on the auditory blend. In the auditory blending procedure, the E says the component sounds (e. g. , m, op) and the S responds with the blended sound (e. g. , mop).

Cycle I teaches the objective behaviors for two bigrams (om and op) and three trigrams (mom, pop, and mop). Table 1 shows the series of steps used to attain this objective. Cycle I begins by teaching three animal sounds to establish the behavior of saying a sound to a graphic image (Level 1 in Table 1). In the next few Cycle I lessons, the child is taught to say the phonic sounds of three letters (m, o, and p) in response to their graphic images (Level 2). Next the child is taught the bigrams om and op on a look-say basis (i. e. , no reference is made to the component letter sounds in Level 3). At Level 4, he is taught to say the component sounds in order, followed by a blended response of both sounds together, as in the previously described model. Level 5 teaches a look-say response to the three trigrams. At Level 6, the child is taught to look at the trigrams (mom, pop, and mop), say the sounds of the first and last two letters, and follow this with the whole-word response. Note that at this point the child can produce the model word-attack behavior in the presence of two particular bigrams and three particular trigrams, but he is not yet expected to generalize this behavior to novel combinations, a task which requires true blending.

Cycle II adds three more single letter-sounds (s, a, t), two more bigrams (at, ot), and three more trigrams (sat, pot, pat). Cycle II omits the look-say responses represented at Levels 3 and 5 in Cycle I, using a previously established blending behavior to teach responses to the new combinations. Finally, Cycle II adds one new skill, word order, to the child's skill repertory.

Cycle III teaches three new single sounds (n, i, f) and exploits the blending behaviors to teach five new bigrams and six new trigrams. Further, Cycle III extends the skills to include four irregular sight words, four punctuation marks, appropriate inflection in oral reading behavior, and capital letters.

After having completed Cycle III, it was expected that the Ss would be able to respond appropriately to all of the content presented in the program. More important than the programmed content, however, it was expected that the child would be able to generalize the programmed skills to correctly analyze and blend novel bigrams and trigrams composed of familiar letters.

SUBJECTS AND PROCEDURE

Initially, Ss were selected from a list of twenty-one children submitted by the school's remedial reading teacher. All of the children on this list were given a laboratory test designed to determine knowledge of any letter sounds, phonic bigrams or trigrams, or simple sight words. Five Ss were selected on the

TABLE 1

CONTENT AND BEHAVIORS TAUGHT IN THE PROGRAM

CYCLE	LEVEL	CONTENT	EXAMPLE		
			Visual Stimulus	Auditory Stimulus	Expected Response
I	1	duck-quack cow -moo dog -woof	Cow Picture	Say the sound... quack	Quack
	2	m, o, p	m	Say the sound.../m/	/m/
	3	om, op	op	Say the sound.../op/	/op/
	4	om, op	om	Say the sound of the first and last letters.../o/, /m/	/o/, /m/
				Say the sound of both together ... /om/	/om/
	5	mom, pop, mop	pop	Read the word... pop	pop
	6	mom, pop, mop	mop	Say the sounds of the first and last two letters.../m/, /op/	/m/, /op/
				Now, read the word... mop	mop
I I	1	s, a, t	s	Say the sound.../s/	/s/
	2	at, ot	at	Say the sounds of the first and last letters.../ae/ /t/	/ae/, /t/
				Say the sound of both together.../at/	/at/
	3	sat, pot, pat	pot	Say the sounds of the first and last two letters.../p/, /ot/	/p/, /ot/
				Now read the word... pot	pot
	4	(word order skill)	mom, sat	Read the words, ... mom, sat	
III	1	n, i, f	f	Say the sound.../f/	/f/
	2	in, on, am, an, it	in	Say the sounds of the first and last letters.../i/, /n/	/i/, /n/
				Now say the sound of both together.../in/	/in/
	3	fat, mat, man, fan, not, sit	fan	Say the sounds of the first and last two letters... /f/, /an/	/f/, /an/
				Now read the word...fan	fan
	4	(sight words) I, is, a, the	the man	Read the words... the man	the man
	5	(names of punctuation marks)period, comma,question mark	sit, man, sit	(E points to comma) What is this mark called?... It's a comma.	"comma"
	6	(oral reading skills)	I am a man	1) (S reads words) 2) (E:" Good, now read it like this," (E reads words with proper inflection). 3) S reads words following E's model.	
	7	(capital letters) M, O, P, S, A, T, N, I, F	F	Say the sound.../f/	/f/

basis of near-zero scores on the test. While testing was in progress, several teachers suggested other children needing remedial help in reading. Seven more Ss were selected from testing results making a total of twelve, seven third-graders and five fifth-grade hold-over students, accepted into the experimental program.

After screening, all Ss were tested on the Wepman Auditory Discrimination Test, Birch Perceptual-Motor Sequencing Test, Birch Audio-Visual Tapping Patterns Test, and the Bender-Gestalt.

A control group of twelve was selected from children who received the laboratory screening test but were not included in the program because they demonstrated a limited sight-word reading vocabulary.

All work with the remedial Ss was conducted in daily sessions (when attendance and schedules allowed) lasting from 3-5 minutes in the early parts of the sequence and from 7-15 minutes in the later parts. In a daily session, the S was seated in the laboratory beside the tutor. The tutor read the lesson or test script showing the appropriate visual stimuli (see examples in Table 1) either on 3 inch by 5 inch index cards or in the pages of a notebook. In some cases, it was necessary to deviate from the lesson script to allow for special problems such as speech difficulties. While sacrificing some of the control necessary for definitive research, this procedure allowed the development of techniques for dealing with these special problems.

Each session was timed with a stop watch, and the tutor noted on the S's record the amount of time used in the session. A rough estimate of the number and nature of the errors occurring in each session was recorded on each S's record.

If an S made more than two errors on one item, or had special difficulty in learning a new element, the tutor repeated the lesson that taught the weak element. If the S did not master the item with this additional practice, a reinforcing activity in the form of a reading game or drill was given in the next session or in several subsequent sessions.

At critical points in the sequence (e.g., at a point between the single-sounds lessons and the bigrams lessons), pre- and posttests were given to determine mastery of previous lessons and to assess knowledge of the next few lessons. These results were used to recycle Ss in lessons not sufficiently mastered and to skip over lessons for which the S already knew the content. This procedure insured 100 percent mastery of previously covered materials and also avoided unnecessary time spent in lessons which the S did not need. Each S continued to receive lessons and tests until 100 percent mastery was attained on the final Cycle III posttest. Two Ss failed to complete Cycle III: one due to an inordinate record of absences from school and the classroom, and the other due to particularly severe memory deficit and attention span problems.

At the end of the school year, a specially designed reading test was administered to all Ss. The test was designed to assess mastery and retention of the programmed content as well as the generalization

of the programmed skills to novel content. The novel content included both meaningful combinations (e.g., mat, top) and abstract combinations (e.g., sop, fam). All combinations were tested on a look-say basis where the child was shown the word and simply instructed to read it, as well as on a blending basis where the child was instructed to say the sounds in the word before being instructed to read it. This dimension was included in the test to assess the degree to which the children were applying the blending model in decoding the words.

RESULTS AND DISCUSSION

The results of the phonics screening pretest are shown in Table 2. The results were compared using the Mann-Whitney U Test. The experimental and control groups performed equally poorly on the single sounds and abstract trigrams, both tasks which involve phonic skills. However, the controls performed significantly better on the meaningful bigrams and trigrams (p < .01 and p < .001 levels respectively). These results show that, although neither group evidenced phonic skills, the control group evidenced a limited sight vocabulary which was not present in the experimental group. It was this factor which excluded the controls from treatment in the program.

Table 3 presents the mean age-level difference achieved by the remedial Ss on the Bender-Gestalt, Birch Perceptual-Motor Sequence, and the Wepman Auditory Discrimination Test compared to the normal population. The Bender-Gestalt and Birch Sequence were designed to identify deficits in visual perception and poor integration of visual perception with motor coordination symptomatic of organic brain injury (1, 9). The Wepman Auditory Discrimination Test identifies children having difficulty perceiving similarities and differences in auditory sequencing (17).

The experimental group fell 3 years below the average 10-year-old child on the Koppitz scoring o

TABLE 2

PERCENTAGE CORRECT IN PHONICS SCREENING PRETEST

	Experimental Group %	Control Group %	Results Mann-Whitney
Single Sounds x = 9	26.9	29.6	U = 52.5 NS
Meaningful Bigrams x = 11	24.2	53.0	U = 25 P = .01
Meaningful Trigrams x = 10	14.2	47.5	U = 20 P = .001
Nonsense Trigrams x = 6	2.8	8.3	U = 48 NS

TABLE 3

MEAN AGE DIFFERENCES FROM NORM IN
ACHIEVEMENT OF REMEDIAL Ss ON
PSYCHOMOTOR TESTS

		Mean Age Level of Remedial Ss	Mean Age Level Scored	Mean Age Level Difference From Norm
	Bender-Gestalt Mean Koppitz Age (N = 12)	10	7	-3
	Wepman Auditory Discrimination (N = 11)	10	below 5	-5
Birch-Perceptual-Motor Sequence	Analysis Items (N = 11)	10	11.5	+1.5
	Synthesis Items (N = 11)	10	7	-3
	Drawing Items (N = 11)	10	below 5.5	-4.5
	Drawing and Matching Items (N = 11)	10	below 5.5	-4.5

he Bender-Gestalt. The mean age-level achieve-
ment of the experimental group on the Birch Sequence
showed differences ranging from -5 years on the
drawing and matching items to +1.5 years on the
analysis items. One S did not take the Birch Sequence.
The test results indicate that some visual perceptual
and integrative deficits were present in the experi-
mental group.

The mean age-level achievement on the Wepman
demonstrated more than a 5 year difference. (One
S was not included due to invalidation of his success.)
An auditory deficit may be indicated, but since the
remedial population all came from Puerto Rican
Spanish-speaking homes, it is impossible to disen-
tangle the effects of oral language experimental dif-
ferences.

Performance of the experimental group is
shown in Table 4, which gives Mean Time per Cycle,
Mean Number of Sessions per Cycle, and Mean Time
per Session. On the average, each child received
about forty-three sessions in a total of about 4 1/2
hours of instructional time during the course of the
experiment. Since the same number of items was
taught in Cycles I and II, the results show that the
children were able to master the Cycle II content in
a little more than half the time required for the com-
parable Cycle I content (92 minutes in Cycle I versus
54 minutes in Cycle II). Since Cycle III included about
twice as much content as the first two cycles and
taught several new skills, it is not possible to com-
pare these aquisition measures to Cycles I and II.

Table 5 indicates mastery of the content as de-
termined by the criterion tests administered during
the program. All Ss achieved 100 percent mastery
in Cycles I and II and in single sound and bigram anal-

ysis acquisition in Cycle III. In Cycle III, 94 per-
cent mastery was obtained for bigrams blending and
92 percent for trigram analysis. Capitals, sight-
words, and punctuation mark labeling and function
were mastered at a 92 percent level. All twelve Ss
completed Cycles I and II. However, since two Ss
did not complete Cycle III, their scores reflected
less than 100 percent mastery of Cycle III content.
Their scores, nevertheless, were included in the
data in Table 5 and account for the less than perfect
mastery obtained for Cycle III.

The results of the end-of-the-year test for pro-
grammed content and skill generalization are shown
in Table 6. These results reflect the mean percent-
age of correct responses to different words tested
regardless of whether the correct response occur-
red as a look-say response, a blended response, or
both. The results of the experimental and control
groups were compared using the Mann-Whitney U
Test. This comparison shows that the experimental
group performed significantly better than the control
group on all sections of the test.

Of particular importance is the significantly bet-
ter performance of the experimental Ss on the "not
programmed" content. Recall that the screening
pretest indicated the controls could read significant-
ly more bigrams and trigrams. However, Table 5
results indicate that after the experimental Ss had
been trained in word-attack skills, they could read
more trigrams than the Control Ss. These results
justify the conclusion that the Ss learned generalized
word-attack skills taught in the experimental pro-
gram.

The degree to which the programmed word-at-
tack skills aided the experimental Ss in decoding the
words tested can be shown in a comparison of the
look-say responses with the blended responses. Re-
call that all content was tested both for a simple
reading response (look-say) and for a reading re-
sponse following an analysis of the word-sounds
(blended). Table 7 shows a comparison of the look-
say and blended responses both for the experimental
and control groups using a Wilcoxon matched-pairs
test.

TABLE 4

EXPERIMENTAL GROUP PERFORMANCE IN
TUTORIAL PROGRAM

	Cycle I	Cycle II	Cycle III
Mean Time in Cycle (Minutes)	92	54	133
Mean Number of Sessions	17	10	16
Mean Time Per Session (Minutes)	5.4	5.4	8.3

TABLE 5

PERCENT OF CONTENT MASTERED BY EXPERIMENTAL GROUP AS DETERMINED ON PROGRAM CRITERION TESTS

| | | Cycle I | | Cycle II | | Cycle III | |
		Pre %	Post %	Pre %	Post %	Pre %	Post %
Single Sounds		24. 8	100	35. 8	100	27. 7	100
Bigrams	Analysis	88	100	92	100	58	100
	Blend	-	100	33	100	46	94
Trigrams	Analysis	58	100	92	100	92	92
	Blend	17	100	36	100	31	100
Word Order		-	-	78	100	-	-
Capitals		-	-	-	-	75	92
Sight-Words		-	-	-	-	44	92
Punctuation		-	-	-	-	25	92

The experimental group performed slightly better on the look-say response for the programmed content. However, this difference did not approach significance. The controls, on the other hand, performed significantly better ($p < .01$) on these words when a simple look-say response was required. These results indicate that asking these Ss to perform a word analysis, a task they had not been trained to do, in-terferred with their ability to read the words.

The effect of the word-attack training becomes apparent when the " not programmed " content is considered. The experimental group performed singificantly better on both meaningful and nonsense words ($p < .01$ and $p < .005$) when an analysis of the word was required before the reading response. The

TABLE 6

PERCENTAGE CORRECT IN POSTTEST

		Test Content	Experimental %	Control %	Mann-Whitney U
Single Sounds (Programmed)		m, o, p, s, a, t, n, i, f	90. 7	41. 6	U = 7 P < .001
Bigrams	Programmed	om, at, on , an, it	93. 3	53. 3	U = 15. 5 P < .001
	Not Programmed	ap, am, af, im, ip	73. 3	16. 7	U = 17. 5 P < .001
Trigrams	Programmed	mom, pop, mop, sat, pat, pot, fat, mat, man, fan, not, sit	91. 3	54. 9	U = 14 P < .001
	Not Programmed Meaningful	mat, top, top, fit, pin, sam	72. 2	40. 3	U = 34 P < .05
	Nonsense	pom, mot, sop, pon, fam, nit	47. 2	13. 9	U = 31 P < .01

TABLE 7

PERCENTAGE OF SELECTED TRIGRAMS READ CORRECTLY UTILIZING A LOOK-SAY APPROACH AND A WORD ANALYSIS AND BLENDING APPROACH

			Look-Say Response %	Blended Response %	Wilcoxen
Experimental	Programmed Meaningful	Trigrams	88.2	82.6	T = 9 N = 8 NS
	Trigrams	Meaningful	44.4	69.4	T = 1 N = 8 P < .01
	Not In Program	Nonsense	18.1	45.8	T = 1 N = 10 P < .005
Control	Programmed Meaningful	Trigrams	54.2	30.6	T = 1 N = 10 P < .01
	Trigrams	Meaningful	37.4	23.6	T = 5 N = 8 NS
	Not In Program	Nonsense	5.6	11.1	No Test

difference is greater when the content is nonsensical than when it is meaningful. The experimental Ss read 15 percent more of the words correctly with the blended response when the content was meaningful and about 30 percent more when the content was nonsensical.

CONCLUSIONS

It should be reemphasized, before stating conclusions, that the purpose of this experiment was to study in detail the acquisition of decoding skills in dyslexics. Since the program involved only an average of 4 1/2 instructional hours per child, and since only a limited set of content and skills was taught, it was not expected that this treatment would have much effect on general reading ability as measured by such instruments as the MRAT. The results of this study, therefore, are in no sense offered as a solution to the problem of reading failure.

Although the meaning of dyslexia is unclear and varies from study to study, there is no doubt that we were dealing with dyslexic children. This conclusion is supported both by their severe reading retardation and by the evidence of perceptual and integrative deficits. The experimental results clearly show that dyslexic children can learn abstract sound-symbol correspondences as well as sight-words. Furthermore, this study provides conclusive evidence that, given appropriate training procedures, dyslexic children can learn to analyze and blend words on a generalized level. No previous studies have documented procedures which would teach these skills to children with pervasive neurological involvement.

Chall, and others (3) raised questions regarding the effect of training on improving auditory blending skills and the effect such improvements might have on general reading ability. The results represented here prove that appropriate exercises can improve auditory blending skills and suggest an effect on general reading ability through the generalized blending performance of our Ss. The results are of importance in view of the finding that the auditory blending deficit is a significant factor in reading disability for neurologically impaired children (8).

We believe that the fundamental key to our success in teaching, analyzing, and blending skills to dyslexic children was the carefully controlled content and skill sequences and monitoring techniques representing an extension of several years of research in the area of decoding (7, 14). The controls necessary to finalize this conclusion, in which the skill and content sequences are not systematically arranged and which use no monitoring techniques, are unfortunately lacking. However, the fact that both experimental and control Ss had experienced 3-5 years of loosely controlled and unmonitored reading instruction prior to this experiment lends some support to this conclusion.

A detailed look at the acquisition records (Sessions, Time, and Criterion Test Score) provides evidence for two interesting observations. First, the finding that Cycle II was mastered in about one-half of the time required for Cycle I lends further support to the conclusion of Gotkin, McSweeney, and Richardson (8:85) that the learning-to-learn phenomenon occurs at an early stage in beginning reading training. Second, the high levels of mastery obtained on the Criterion Test represent a typical result with programmed instruction. The usual bell-shaped achievement distribution yields to a clustering of scores around a high level of achievement. The high mastery results from the programmed monitoring techniques and reflects the success of applying these techniques with a dyslexic population.

The evidence collected here on the generalized blending skill adds to the evidence already cited (7, 10, 14) that children can be effectively trained in applying this skill. However, this study represents an extension of these earlier results to dyslexics as well as suggesting other insights into the nature of this skill. It was shown that performing a visual analysis interfered with the control Ss' ability to read the words tested. The implications of this result become apparent when one considers that these Ss demonstrated some minimal phonic skills (e. g. , initial sounds) while the application of these skills actually interferes with his ability to decode words. The interesting question implied by this is: How much and what kind of training is necessary to transform these into useful decoding skills? Another interesting insight into the blending skill is provided by the evidence that it is applied differentially to meaningful and nonsense words. On the one hand, it is clear that the skill led to the successful decoding of a greater percentage of novel meaningful words than nonsense words. On the other hand, when the visual analysis was a required part of the response, there was a greater increase in the percentage of nonsense words correctly decoded than meaningful words. One possible interpretation of these results is that children were using their blending skills when simply asked to read the word (the " look-say" response) but were more successful in applying it to words already in their vocabulary (meaningful). However, being specifically instructed to analyze the words orally (blended response) significantly increased the number of words they could decode for both categories. It would follow, then, that there would be a greater increase for the nonsense words since several of the meaningful words had already been successfully blended when the look-say response was required.

These data raise questions of how decoding skills are actually used in reading. Do they become highly integrated processes that occur at a high rate of speed? Do they merely serve to help the reader decode an unfamiliar word until that word becomes integrated into his sight vocabulary? These data suggest different ability levels in the application of the blending skill. When does this skill emerge? To what degree of utility can this skill be attained? These questions, and a host of other relevant questions, merely reflect the primitive state of our knowledge of the process of decoding.

Finally, although the work reported here is limited, the initial success with this program would indicate that a highly-structured, individually tutored, programmed approach may be a promising avenue of investigation leading to real solutions to the dyslexic problem. Such solutions would take several more years of developmental work extending these techniques to other decoding skills and to comprehensive skills.

REFERENCES

1. Birch, H. G. ; Lefford, A. , " Visual Differentiation Intersensory Integration, and Voluntary Motor Control," Child Development Monograph, 32: No. 2, 1967.

2. Chall, Jeanne, Learning To Read: The Great Debate, McGraw-Hill, New York, 1967, p. 103.

3. Chall, Jeanne; Roswell, Florence; Blumenthal, Susan, H. , " Auditory Blending Ability: A Factor in Success in Beginning Reading," The Reading Teacher, 17:113-118, 1963.

4. Delacato, Carl, H. , The Diagnosis and Treatment of Speech and Reading Problems, Charles C. Thomas, Springfield, Illinois, 1964.

5. de Hirsch, Katrina; Jansky, J. J. ; Langford, Wm. S. , Predicting Reading Failure, Harper and Row, New York, 1966.

6. Ellson, D. G. ; Harris, Phillip; Barber, Larry, " A Field Test of Programmed and Directed Tutoring," Reading Research Quarterly, 3:307-367, Spring 1968.

7. Gotkin, Lassar, G. ; McSweeney, Joseph; Richardson, Ellis, The Development of a Beginning Reading Program, Institute for Developmental Studies, New York University, New York, June 1969.

8. Huset, Martha, " Relationship between Difficulty in Auditory Blending and Some Diagnostic Indicators of Organicity in Children of Average or Superior Intelligence with Reading Disability", unpublished master's thesis, The City College of New York, New York, June 1961.

9. Koppitz, Elizabeth, M. , Bender-Gestalt, Test for Young Children, Grune and Stratton, New York, 1964.

10. McSweeney, Joseph F. ; Richardson, Ellis, The Application of Process-Oriented Curricula to Beginning Reading Instruction, New Jersey Early Childhood Learning Development Center, Newark, July 1969, p. 28.

11. Money, John (ed.), Reading Disability, Johns Hopkins, Baltimore, 1966, p. 74.

12. New York Times, 1:3, February 15, 1970.

13. Rawson, Margaret B., Bibliography on the Nature, Recognition, and Treatment of Language Difficulties, The Orton Society, 1966, pp. 57-71.

14. Silberman, Harry F. , Reading and Related Verbal Learning, System Development Corporation, Santa Monica, California, 1963.

15. Silver, Archie A. ; Hagin, Rosa A. , Strategies of Intervention in the Spectrum of Defects in Specific Reading Disability, Bulletin of the Orton Society, 17:43-44, 1967.

16. Wasserman, Miriam, "Planting Pansies on the Roof, " The Urban Review, 3:30-34, January 1969.

17. Wepman, Joseph M. , Auditory Discrimination Test, University of Chicago, Chicago, Illinois, 1958.

18. Woolman, Myron, " Cultural Asynchrony and Contingency in Learning Disorders, " Learning Disorders, Vol. 1, U. S. Government Printing Office, 1965.

SEX, GRADE LEVEL, AND RISK TAKING

ON OBJECTIVE EXAMINATIONS

MALCOLM J. SLAKTER, ROGER A. KOEHLER, SANDRA H. HAMPTON
State University of New York at Buffalo

ROBERT L. GRENNELL
State University College at Fredonia

ABSTRACT

Students in one school system in grades 5 through 11(522 boys, 548 girls) responded to an objective examination which incorporated a measure of risk taking. The study was replicated in a second school system (600 boys, 691 girls). In each case the proportion of risk-taking variance associated with variation in grade level was approximately .10 (significant at the .05 level), with higher risk in grades 5, 6, and 7 than in grades 8, 9, 10, and 11. Boys took greater risks than girls in both school systems, but the proportion of risk-taking variance explained by sex was low (approximately .01) and significant (at the .05 level) in only one school system. There was no interaction between grade level and sex.

ALTHOUGH risk taking is becoming an increasingly interesting variable in educational and psychological research, comparatively little is known about its relation to age or sex. For example, Wallach and Kogan (13) found that older Ss (mean age approximately 70) were more conservative than college students on a hypothetical choice-dilemmas instrument. In studies of 6- to 10-year-old children, Kass (4) found no age difference in gambling with pennies in a slot machine. However, Cohen (2: chapter 5) in a risk-taking situation with candy for prizes, found that 9-year-old children took greater risks than 12-year-olds, who in turn took greater risks than 15-year-olds.

With respect to sex, Wallach and Kogan (12, 13) have found no consistent sex differences in risk, and little evidence of feminine conservativeness. Kass (4), however, found boys selected greater risks than girls with the slot machines. On the other hand, using a decision-making task with candy as the prize, Slovic (10) found a sex-by-age interaction; i.e., no sex difference in younger children (ages 6-10), but with older children (ages 11-16) greater risk was manifested by boys.

Risk taking on objective examinations (RTOOE) is defined as guessing when the examinee is aware that there is a penalty for incorrect responses (6).

A previous study (7) noted (a) the potential usefulness of RTOOE to psychologists as a disguised measure of risk taking, and (b) the effect of RTOOE on test score, which merits the attention of individuals concerned with educational measurement. Specifically, with RTOOE, no sex differences were found in college students (6, 8, 9). However, eighth-grade females were found to be greater risk takers than eighth-grade males (7), while the opposite was found for ninth-grade students (11). The purpose of the present study, therefore, was to (a) devise measures of RTOOE that would be appropriate for use in grades 5 through 11, and (b) administer the measures to suitable grades in order to observe the relation of RTOOE with grade level or sex.

METHOD

The measures of RTOOE were based upon the use of nonsense items, where a nonsense item is defined as one that has no correct (or best) answer, and no incorrect answer for the given population. Previous research (6, 7, 8, 9) has suggested that five nonsense items embedded in five legitimate items would provide suitable test characteristics. In addition, since there is evidence that RTOOE is a general trait across different types of examinations (7), convenient synonym-antonym vocabulary items were the type employed in the measures. Subjects were

directed to indicate whether the words had the same or opposite meaning, and were informed of the penalty for incorrect responses. The following is an example of a nonsense item used in the measures:

7. marnel............mild

Since "marnel" is meaningless, the item has no answer. Hence, any response (i. e., "same" or "opposite") is assumed to be an example of RTOOE behavior; if the item is omitted, a lack of RTOOE behavior is indicated.

In order for grade trends to be examined, the nonsense items which formed the basis of the risk measures were constructed so that they could be used at all grade levels; the legitimate items were selected to be appropriate for the particular grade level, and generally appeared at a single grade level. The RTOOE score assigned to an S was the proportion of nonsense items attempted. The instruments to measure RTOOE were constructed to the above specifications, and were tried out on selected classes from grades 5 through 11 in a large village in western New York State. The procedure used for estimating reliability was the Kuder-Richardson formula 20 (K-R 20), which is a measure of internal consistency, or the tendency of the test items to be homogeneous (e. g., 3, 161). The analysis of these preliminary data revealed a median K-R 20 of .78 across the seven grades for the RTOOE measure. Additional evidence of the reliability and construct validity of this method of measuring RTOOE is provided elsewhere (7).

Subjects for the study were all available public school students in grades 5 through 11 in the same New York village where the preliminary data were collected. There were a total of 1,070 Ss, consisting of 522 males and 548 females. The number in each grade ranged from 118 to 228. The entire study was then replicated in a small city in northern Michigan, with a total of 1,291 Ss, consisting of 600 males and 691 females. The number in each grade for the Michigan study ranged from 140 to 208. The tests were administered to the Ss in their own classrooms by their own teachers. The teachers had been previously instructed as to standardized procedures of administration. The Ss were generally led to believe that they were taking another aptitude examination in their school's testing program. The tests were given as Part I of an "aptitude" examination on the same day to all classes in a given school, and within several days to the entire set of classes in a school system.

RESULTS AND DISCUSSION

The K-R 20 reliabilities for the RTOOE measure are presented in Table 1. The values for the seven grades in the New York system ranged from .68 to .86, with a median value of .83. With the Michigan Ss, the reliabilities ranged from .78 to .88 with a median of .86. Hence, it appears that the RTOOE measure was generally reliable across grades 5 through 11 for both school systems.

Table 2 presents the mean risk scores for the New York State Ss by sex and grade. Numbers in parentheses are the sample sizes; the estimate of mean square within and its corresponding degrees of freedom are provided at the bottom of the table. A sex-by-grade factorial analysis of variance was performed; because the cell frequencies were unequal, an exact least squares analysis proposed by Bock (1) was utilized. The results indicated that the grade effects (with the effects of sex eliminated) were significant at the .05 level. The proportion of RTOOE variance associated with variation in grade level or η^2 (5) was approximately .11. Neither the sex effects (with grade effects eliminated) nor the sex-by-grade interaction effects (with both main effects eliminated) were significant at the .05 level.

The results for the Michigan replication are provided in Table 3. Once again, the grade effects (with the sex effects eliminated) were significant at the .05 level, with a corresponding η^2 of approximately .10. In addition, the sex effects (with the grade effects eliminated) were significant at the .05 level. However, η^2 for the latter was approximately .01. As before, the sex-by-grade interaction effects (with both main effects eliminated) were not significant at the .05 level.

It would appear, therefore, that there is evidence that grade level (or age) and RTOOE are related, and that the relation is fairly sizable (approximately 10 percent of the variance in RTOOE is accounted for by differences in grade level). From an inspection of the data, it would further appear that Ss at grade levels 5 through 7 have greater mean RTOOE than Ss at grade levels 8 through 11. Scheffe post hoc comparisons confirm this conjecture at the .05 level, for both the New York and Michigan data. The finding that risk taking was greater for younger children is consistent with the results of Cohen (2).

The conclusions with respect to sex are not as clear as those with respect to grade level. There is some evidence with the Michigan data that sex is related to RTOOE (.05 level), with the data indicating that males are greater risk takers. However, the relation between sex and RTOOE for the Michigan data was not strong (approximately 1% of the variance in RTOOE can be accounted for by sex), and the relation was not replicated with the New York data. Hence, the findings with respect to sex are somewhat similar to that of Wallach and Kogan (12, 13) in that they are not consistent. At the same time, the findings somewhat support the results of Kass (4) in that males were greater risk takers than females in one of the school systems. It is, however, interesting to note that the present findings are in direct opposition to a previous study (7) which resulted in eighth grade females demonstrating consistently higher RTOOE

TABLE 1

K-R 20 RELIABILITIES FOR GRADES 5 THROUGH 11

Ss \ grade	5	6	7	8	9	10	11
New York	.78	.68	.85	.76	.86	.85	.83
Michigan	.88	.86	.87	.78	.84	.87	.85

TABLE 2

MEAN RISK SCORE FOR NEW YORK STATE Ss (SAMPLE SIZE IN PARENTHESES)

grade sex	5	6	7	8	9	10	11	
M	.92 (75)	.92 (63)	.93 (59)	.79 (48)	.72 (99)	.76 (85)	.61 (93)	.79
F	.85 (60)	.87 (56)	.96 (71)	.73 (70)	.71 (129)	.68 (80)	.66 (82)	.76
	.89	.90	.95	.75	.72	.72	.63	

MS_w = .0945, d.f. = 1056

than eighth grade males. Hence, it seems that the relation between sex and risk (or more specifically between sex and RTOOE) remains unclear.

It should be noted that grade level differences in mean RTOOE are confounded not only with differences in Ss (since these are cross sectional studies) but also with differences in teachers or test administrations. The teacher influence could certainly be potent since in their classes certain teachers might advise their students to " guess at everything " or to " omit questions unless you are sure of the answer," etc. The administrator effect would appear if, on the criterion examination, certain teachers strayed inadvertently from the standardized procedure.

In order to investigate the possibility of a teacher or administrator effect, an expanded analysis of the present data was performed. The expanded analysis considered classes as nested within grades (Ss who shared the same teachers or took the criterion examination together were considered as one class). In addition, both grade and class-within-grade were crossed with sex. For the New York Ss, only the grade effects were significant at the .05 level, with the corresponding η^2 approximately equal to .11. Hence, there was no evidence that classes or administrations within grades were an important source of variation.

With the Michigan data, however, both the sex-by-grade and the sex-by-class-within-grade interactions (each with all other effects eliminated) were significant at the .05 level. For each interaction, the η^2 was about .01. Classes nested within grades (with all main effects eliminated) was significant at the .05 level, with η^2 approximately equal to .12. Hence the results from the Michigan data appear to be much more complicated than those from the New York data, with some evidence for a teacher or administrator effect, but also with some weak interactions; e.g., the sex effect varied from class to class within a given grade. A possible explanation for the more complicated Michigan data is the much shorter time period spent with those teachers in their orientation session; i.e., the strong relation between classes within grades and RTOOE is perhaps simply reflective of a lack of success in accomplishing standardization in the administration of the tests. In addition, the two significant interactions were both weak, and one (sex-by-class-within-grade) difficult to explain. (A plausible conjecture that the sex of the teacher was involved proved to be groundless.)

In summary, there is evidence that RTOOE is related to grade level (or age), with higher RTOOE associated with grades 5, 6, 7 than with grades 8, 9, 10, 11. If there is a relation between RTOOE and sex,

TABLE 3

MEAN RISK SCORE FOR MICHIGAN Ss (SAMPLE SIZE IN PARENTHESES)

grade sex	5	6	7	8	9	10	11	
M	.71 (83)	.83 (63)	.82 (93)	.51 (104)	.64 (80)	.55 (95)	.71 (82)	.67
F	.61 (92)	.73 (77)	.78 (88)	.41 (104)	.64 (100)	.42 (103)	.65 (125)	.60
	.66	.77	.80	.46	.64	.48	.68	

MS_w = .1340, d.f. = 1275

it appears to be weak. Finally, there is some preliminary evidence that the teacher or test administrator may have some effect on RTOOE.

With respect to education and/or psychology, we may consider the following implications. First, it would appear that students become more conservative on RTOOE as they grow older. Whether this increased conservativeness is due to maturation, the educational process, or combinations of these and other factors is not known. Since, the present study was cross-sectional, one might speculate that the mean differences were due to a shifting population. For example, school dropouts might be high in RTOOE, and their disappearance from the school scene would result in lower RTOOE. However, this phenomenon would be accompanied by a decrease in RTOOE variance. An investigation of the RTOOE variance across grades revealed either a steady pattern, or an increase, but not a decrease. Hence, the dropout conjecture seems less plausible. Planned longitudinal studies should shed more light on the shifting population hypothesis.

Second, since there is evidence that RTOOE affects aptitude or achievement scores, test constructors should be aware that (a) there may be an administrator effect on RTOOE, (b) lower RTOOE is more characteristic of grades 8 through 11 than grades 5 through 7, (c) there may be school or geographic differences in RTOOE (note the mean differences between the New York and Michigan school systems), and (d) the mean difference in RTOOE for boys and girls appears to be small.

Lastly, there is evidence that reliable measures of RTOOE can be constructed for students as early as the fifth grade.

FOOTNOTE

1. This research was supported by the Teacher Education Research Center, State University College at Fredonia, New York.

REFERENCES

1. Bock, R. Darrell, " Programming Univariate and Multivariate Analysis of Variance," Technometrics, 5:95-117, February 1963.

2. Cohen, J., Chance, Skill, and Luck, Penguin, Baltimore, Maryland, 1960.

3. Cronbach, Lee J., Essentials of Psychological Testing, Harper and Row, New York, 1970, 752pp.

4. Kass, Norman, " Risk in Decision Making as a Function of Age, Sex, and Probability Preference," Child Development, 35:577-582, 1964.

5. McNemar, Quinn, Psychological Statistics, John Wiley and Sons, Inc., New York, 1969, 529 pp.

6. Slakter, Malcolm J., " Risk Taking on Objective Examinations," American Educational Research Journal, 4: 31-43, January 1967.

7. Slakter, Malcolm J., " Generality of Risk Taking on Objective Examinations," Educational and Psychological Measurement, 29: 115-128, 1969.

8. Slakter, Malcolm J.; Cramer, Stanley H., "Risk Taking and Vocational or Curriculum Choice," Vocational Guidance Quarterly, 18: 127-132, December 1969.

9. Slakter, Malcolm J.; Koehler, Roger A., " A New Measure of Risk Taking on Objective Examinations," California Journal of Educational Research, 19: 132-137, 1968.

10. Slovic, Paul, " Risk Taking in Children: Age and Sex Differences," Child Development, 37: 169-176, March 1966.

11. Swineford, Frances, " Analysis of a Personality Trait," Journal of Educational Psychology, 32: 438-444, 1941.

12. Wallach, Michael A.; Kogan, Nathan, " Sex Differences and Judgment Processes," Journal of Personality, 27: 555-564, 1959.

13. Wallach, Michael A.; Kogan, Nathan, " Aspects of Judgment and Decision-Making; Interrelationships and Changes with Age," Behavioral Science, 6: 23-36, 1961.

EFFECTS OF ETHNIC BACKGROUND, RESPONSE OPTION, TASK COMPLEXITY AND SEX ON INFORMATION PROCESSING IN CONCEPT ATTAINMENT

GLENN E. TAGATZ, LEE R. HESS, JANE A. LAYMAN
Marquette University

DEAN GARRISON
Mesa Alta School, Bloomfield, New Mexico

ABSTRACT

The effects of (1) Navaho, Spanish, and Anglo-American ethnic backgrounds, (2) response options of inclusion, exclusion, and indeterminate categorization, (3) two degrees of stimulus complexity, and (4) sex were examined during information processing in concept attainment. Subjects were a stratified sample of eighty four ninth grade students selected from all the ninth grade students attending a Bloomfield, New Mexico, municipal school. The dependent variable consisted of responses to the Tagatz Information Processing Test (TIPT). A 3x3x2x2 repeated measures analysis of variance yielded four significant sources of variance. The hierarchial mean performances were each significantly different from the other. The ethnic group by response option interaction indicated that the hierarchy of Navaho, Spanish, and Anglo-American was not evident for items of indeterminate categorization. For items of this type, the means of the three groups were not statistically different from each other. The response of stimulus complexity interaction indicated that items in the complex presentation with an exclusion response were significantly different from all other cells in the interaction.

THE EFFECTS of a person's genetic and cultural backgrounds upon his IQ score and academic performance has received considerable attention in recent literature (1, 2, 3, 4, 5). Jensen (3) has hypothesized that 80 percent of the IQ variance of European and North American populations is determined by heritability (H). Elkind (2), on the other hand, interprets differences in terms of Piaget's Structuralism, believing that intelligence is developed through experience.

The examination of the pro and con arguments indicates many areas of agreement between Jensen and his critics. The foremost is that members of different races, cultures, and socioeconomic strata (SES) do possess differing intellectual abilities. The cause can be genetic, psychogenic, cultural, or a combination of these factors, and most schools as they presently operate, appear to maximize rather than minimize these differences (4). Another substantial area of agreement is that more study is needed to determine such factors as (1) the cause of the large differences among racial groups in school performance and test scores, (2) dysgenic trends, and (3) the best method(s) to educate culturally distinct groups.

In a comparison of information processing during concept attainment using third and fourth graders within the same school, Tagatz, Layman, and Needham (9), found no significant differences between (SES). Marascuilo and Amster (8) did find a significant difference favoring the upper SES over lower SES children. Their Ss were fifth and sixth grade children in separate schools. The stimuli in the present study are similar to those used by Tagatz, Layman, and Needham (9) where SES factors were minimal or nonexistent. Jensen (3), summarizing several of his own research articles, found that with certain learning tasks lower-class children be they white, Negro, or Spanish-American, perform as well as middle-class children in the same IQ

range. Disparate results such as these illustrate the need for research about the nature of systematic differences between sociocultural groups.

THE PROBLEM

The general purpose of this study was to investigate differences in information processing during concept attainment of ninth grade students from Navaho, Spanish, and Anglo-American ethnic backgrounds. The following specific question was examined:

What are the effects of the following variables on information processing during concept attainment: (a) ethnic background: Navaho, Spanish, and Anglo-American; (b) response option: inclusion, exclusion, and indeterminate categorization; (c) task complexity: two degrees of stimulus complexity; and (d) sex.

METHOD

Subjects

Subjects for the study were a stratified sample of eighty-four ninth grade students selected from all the ninth grade students attending a Bloomfield, New Mexico, municipal school. The Ss were stratified as follows: fourteen Anglo males, fourteen Anglo females, fourteen Navaho males, fourteen Navaho females, fourteen Spanish-American males, and fourteen Spanish-American females. The mean age was 16 years, 4 months.

The Ss represented a wide range of ability levels. The mean IQ score on the Otis Quick Scoring Test of Mental Ability was 96.67 with a standard deviation of 16.12. The Anglo males had a mean IQ of 98.74, Anglo females 110.24, Navaho males 80.44, Navaho females 80.60, Spanish-American males 95.89, and Spanish-American females 99.86. In summary, the ethnic groups considered represented a wide range of intellectual abilities.

Experimental Materials

The TIPT has been described by Tagatz and Meinke (10) and by Lemke, Klausmeier, and Harris (6). The TIPT contains sixty items divided into two subtests. The first subtest consists of thirty items in which one instance, either an exemplar or non-exemplar, is presented with an exemplar focus instance. The task is to specify the inclusion, exclusion, or indeterminate membership of another instance to membership in a group of instances exemplifying a concept. Fifteen exemplar and fifteen non-exemplar items were used. In the exemplar items, ten test instances—i. e., instance for which membership was to be determined—were definitely exemplars of the same concepts that the focus instance exemplified. The membership of the remaining five test instances could not be determined. In items presenting a focus and a non-exemplar instance, ten test instances were definitely non-exemplar and five again were indeterminate. Thus, the first sub-test of thirty items could be scored on the basis of information presented—fifteen exemplars and fifteen non-exemplars—or on the basis of membership—ten exemplars, ten non-exemplars, and ten instances of indeterminate membership. These three conditions of membership constituted the response options available to Ss.

The second subtest of thirty items was constructed with the use of the same focus and test instances as the first. The information presented in addition to the exemplar focus instance consisted of two other instances, rather than one as in the first subtest. One of the two instances for each item was an additional exemplar; the other was the same in kind as its counterpart in the first subtest. The answers to items of Subtest 2 were exactly the same as Subtest 1.

PROCEDURES

In order to ensure that the directions and examples were fully understood, the TIPT was administered over three class periods. The first was used for orientation. Transparencies of the instructions were prepared and used in explaining each part of the instructions. Subjects were encouraged to stop the examiner at any point for clarification. Sample problems were worked and explained. At the second meeting, students silently reread the instructions, filled out the answer sheets, and used as much time as necessary to complete Subtest 1. The third period was used to examine the instructions for Subtest 2. As with Subtest 1, no time restrictions were imposed.

ANALYSIS AND RESULTS

Responses to this test constituted the dependent variable for the study, and a 3x3x2x2 analysis of variance was performed on these data. Two factors were sex and ethnic groups; two factors consisted of repeated measures inherent in the construction of the test, namely the degree of stimulus complexity and the inclusion, exclusion, and indeterminate categorization of the response option.

Of the four variables studied, and their interactions, four sources of variance were found to be statistically significant (see Table 1). These were the effects of (1) the three ethnic groups; (2) the three response options inclusion, exclusion, and indeterminate categorization; (3) the ethnic group by response option interaction; and (4) the response option by stimulus complexity interaction.

The mean performances of the Navaho, Spanish, and Anglo-Americans were 25.75, 30.54, and 34.21 respectively. Duncan's Multiple Range Test indicated that each mean was significantly different from all others.

The mean performances of the inclusion, exclusion, and indeterminate response options were 12.95, 7.89, and 9.24 respectively. Here too Duncan's Multiple Range Test indicated that each mean was significantly different from the other two.

In Table 2 are reported the means of the performance scores for the ethnic group by response option interaction and Duncan's Range for these data. The hierarchy of Navaho, Spanish, and Anglo-Americans was most apparent in inclusion—type items with

TABLE 1

ANALYSIS OF VARIANCE OF INFORMATION
PROCESSING SCORES

Source	df	MS	F
Sex (S)	1	19.44	1.38
Ethnic Group (E)	2	83.80	5.95**
S x E	2	21.79	1.55
Ss w groups	78	14.08	
Response Option (R)	2	287.65	42.94**
S x R	2	.58	
E x R	4	19.19	2.86*
S x E x R	4	3.81	
R x Ss w groups	156	6.69	
Stimulus Complexity (SC)	1	8.90	2.24
S x SC	1	6.45	1.62
E x SC	2	1.23	
S x E x SC	2	1.66	
SC x Ss w groups	78	3.97	
R x SC	2	12.18	3.25*
S x R x SC	2	.71	
E x R x SC	4	.30	
S x E x R x SC	4	4.95	1.32
R x SC x Ss w groups	156	3.75	

* $p < .05$
** $p < .01$

each cell significantly different from every other
cell. With items of indeterminate categorization,
the hierarchy was only partially evident in that the
Spanish and Anglo-Americans were significantly dif-
ferent from the Navaho's but not different from each
other. For items with responses of exclusion, the
differences between the groups were not significant.

The means of the performance scores for the re-
sponse option by stimulus complexity interaction and
Duncan's Multiple Range Test for these data are pre-
sented in Table 3. Here the complexity created by
the addition of another exemplar in each item inter-
fered with performance on exclusion-type response
items. In the complex presentation, two exemplars
and one nonexemplar were shown, and it seems that
this additional exemplar was given attention, to the
detriment of the non-exemplar processing. Items

TABLE 2

MEANS OF PERFORMANCE SCORES FOR THE
ETHNIC GROUP BY RESPONSE OPTION
INTERACTION AND DUNCAN'S NEW MULTIPLE
RANGE TEST

	Inclusion	Exclusion	Indeterminate	Duncan's Range
Anglo	15.36	8.35	10.50	1.57
Spanish	12.89	7.50	9.96	
Navaho	10.61	7.82	7.32	

TABLE 3

MEANS OF PERFORMANCE SCORES FOR THE
RESPONSE OPTION BY STIMULUS COMPLEXITY
INTERACTION AND DUNCAN'S NEW MULTIPLE
RANGE TEST

	Inclusion	Exclusion	Indeterminate	Duncan's Range
Single Presen- tation	6.39	4.38	4.68	.67
Complex Presen- tation	6.56	3.51	4.58	

in the complex presentation with an exclusion re-
sponse were significantly different from all other
cells in the interaction.

DISCUSSION

The purpose of this study was to investigate dif-
ferences in cognitive functioning of adolescents from
the three ethnic backgrounds, but not to determine
whether heredity or environment is the cause for such
differences. While this latter question has academic
interest, the task of schools is to educate all citizens
optimally; and, as such, functional differences are
of practical as well as theoretical importance.

The conclusion that is reached from the data is
that performance differences exist among these three
ethnic groups. Beyond the total quantitative differ-
ences observable among the three groups, differences
were found among groups in the degree of success
with which various kinds of information were pro-
cessed. The response option by ethnic group inter-
action revealed marked differences in the ability of
the groups to deal with material exemplary of a con-
cept. A similar trend was found when the material
was indeterminate in its ability to yield relevant in-
formation about the attributes of a concept. When the
stimulus was designed to specify the exclusion of an
instance to a conceptual category, the differences
among the three groups were not evident.

It has been pointed out by Lemke, Tagatz, and
Meinke (7) that concept attainment of exemplar in-
formation correlates highly with all curricular fac-
tors. This was not the case with processing of non-
exemplar information. Most cognitive learning is
based on illustrating concepts, and less attention is
given to differences between conceptual categories.
This suggests that a learning set may be operative
in much human learning similar to that found by Ta-
gatz, Walsh, and Layman (11), where Ss initially
left to their own devices did not adapt to the exigen-
cies demanded of them in information processing of
both exemplar and non-exemplar types (conservative
instructions). It is further suggested that, when pro-
cessing a novel type of information (exclusion type
items), differences among the three groups are not

as marked or are perhaps even nonexistent. This may have resulted because experiential differences did not influence performance for this novel task, or it may have been that all Ss were equally unclear as to what they were supposed to do. The fact that all groups scored above chance on these items partially negates the second of these alternatives. Performance differences on the novel exclusion type items were not evident, even though the Ss represented such a wide range of ability levels. This explanation supports Elkind's conception of intelligence rather than Jensen's.

The response option by stimulus complexity interaction further substantiates the importance of exemplar information processing as a normal mode of human functioning. When an additional exemplar was presented, performance deteriorated markedly for the exclusion type items. Subjects have either an inherent or learned preference to deal with positive information. It is a cognitive focusing on figure to the exclusion of ground so often seen in other perceptual tasks. It is important to note, though, that in complex cognitive functioning essential information is gleaned by comparing instances (exemplars or non-exemplars), and it seems our schools may be remiss in properly preparing students for such practices.

REFERENCES

1. Cronbach, L. J., " Heredity, Environment, and Educational Policy," Harvard Educational Review, 39:338-347, 1969.

2. Elkind, D., " Piagetian and Psychometric Conceptions of Intelligence," Harvard Educational Review, 39: 319-337, 1969.

3. Jensen, A. R., " How Much Can We Boost IQ and Scholastic Achievement ?" Harvard Educational Review, 39: 1-123, 1969.

4. Jensen, A. R., " Reducing the Heredity-Environment Uncertainty: A Reply," Harvard Educational Review, 39: 449-483, 1969.

5. Jensen, A. R., " Social Class, Race, and Genetics: Implications for Education," American Educational Resource Journal, 5: 1-42, 1968.

6. Lemke, E. A.; Klausmeier, H. J.; Harris, C. W., " Relationship of Selected Cognitive Abilities to Concept Attainment and Information Processing," Journal of Educational Psychology, 58: 27-35, 1967.

7. Lemke, E. A.; Tagatz, G. E.; Meinke, D. L., " The Relationship Between Conceptual Learning and Curricular Achievement," Journal of Experimental Education, 38: 70-75, 1969.

8. Marascuilo, L.; Amster, H., " The Effect of Variety in Children's Concept Learning," California Journal of Educational Resources, 17: 113-125, 1966.

9. Tagatz, G. E.; Layman, J. A.; Needham, J. R., " Information Processing of Third and Fourth Grade Children," Contemporary Education, in press.

10. Tagatz, G. E.; Meinke, D. L., " Information Processing in Concept Attainment Tasks," Teacher College Journal, 37: 182-186, 1966.

11. Tagatz, G. E.; Walsh, M. R., Layman, J. A., " Learning Set and Strategy Interaction in Concept Learning," Journal of Educational Psychology, 60: 488-93, 1969.

BOOK REVIEWS

contrasting, of searching for "meaning" in the shape of general principles or implications. (2) The student must get excited about the subject of study. (3) The topic of study must touch the student's personal experience, personal problems, or personal ideals. (4) A new type of examination should be used which will not be based upon factual knowledge alone.

One of the principal features of the book is the discussion of "syndicate methods." By syndicate the author means the leadership of small class sections in which the instructor keeps rather close contact, but the group plans their own learning activities.

A very stimulating discussion is done in Chapter 5, "The Contribution of Practical Experience." Among the major topics discussed by the author is that of the discrimination of values giving justice as he does as to how the building up of values in the minds of students might be stimulated and assisted by the instructor.

Harl R. Douglass, Reviewer
Lecturer and Consultant
4185 Pinon, Boulder, Colorado

ESSENTIALS OF PSYCHOLOGICAL TESTING (Third Edition)

Cronbach, Lee J. (New York: Harper and Row, 1970), pp. xxx + 752.

THIS HARDY perennial needs no introduction to most readers. The popular first and second editions have established the text as a reliable source of information for both students and practitioners. What, then, could

THE EFFICACY OF PLAYING HARD-TO-GET[1]

ELAINE WALSTER, G. WILLIAM WALSTER
The University of Wisconsin

ELLEN BERSCHEID
University of Minnesota

ABSTRACT

It has often been suggested that individuals will prefer dates who play "hard-to-get." Two experiments were conducted to test the hypothesis that teen-agers will assume that a hard-to-get individual is more social-ly desirable than a person whose high regard is easily obtained. This hypothesis was not confirmed; the results were opposite to those predicted. It appears that playing hard-to-get is not an effective strategy for increasing one's status. Apparently, all the world does love a lover.

FIFTEEN YEARS ago, teachers were clear about their assignment; the formulation of behavior-al objectives was restricted to the mastery of the three R's. Today, teachers express great concern about their students' social adjustment, peer rela-tions, self-concepts, mental health, and moral and personality development. In fact, many educators argue that such adjustments are prerequisite for ac-ademic achievement. It is not uncommon to find that goals concerned with social and personality adjust-ment take priority to goals concerned with concept attainment in many experimental or innovative pro-grams.

At the same time, there are very few guidelines provided for the teacher who wishes to contend with the emotional needs of his students. Neither teach-er trainees' preparatory course-work nor the re-search literature provides such information. Thus, even the conscientious teacher is forced to rely en-tirely on his own intuition in these areas.

The following study investigates some of the fac-tors which influence teen-agers' perception of one another's social worth.

THE ELUSIVE PERSON

Socrates' advice to Theodota, a hetaera, on how to win friends and influence people was direct: play "hard-to-get":

. . . They will appreciate your favors most highly if you wait till they ask for them. The sweetest meats, you see, if served before they are wanted seem sour, and to those who had enough they are positively nauseating; but even poor fare is very welcome when offered to a hungry man. (Theodota inquires) and how can I make them hunger for my fare? (Socrates' reply) Why, in the first place, you must not of-fer it to them when they have had enough—but prompt them by behaving as a model of propri-ety, be a show of reluctance to yield, and by holding back until they are as keen as can be; for then the same gifts are much more to the recipient than when they are offered before they are desired (8).

The notion that an individual can become desirable by playing hard-to-get is not only part of our folk-lore but part of the folklore of other times and coun-tries. While Ovid, the Kama Sutra, and Dear Abby all agree that the lover should not display his affec-tion too readily, no experimental evidence exists to document the effectiveness of the hard-to-get strat-egy.

There are some correlational data which indicate that those who appear to be greatly in need of affec-tion are not held in high regard. Ehrlich (personal communication, 1969) found that mental patients who admitted possessing a strong need for approval were less popular among other patients and among the staff than were other patients. Ehrlich points out that her results agree with those reported by Crowne

and Marlowe (1) who found a negative correlation between the "approval dependence" of fraternity men and their popularity with other men. In these correlational studies it is not possible to determine if individuals have a strong need for approval because they have been rejected by others or if their desperate need for approval causes them to be rejected.

If being "hard-to-get" does in fact increase one's desirability, several theories might account for this phenomenon.

1. Dissonance Theory (3): The person who is hard-to-get requires a suitor to expend more effort in her pursuit than he would normally expend. One way the suitor can justify his unwarranted expenditure of energy is by aggrandizing the hard-to-get woman.

2. Learning Theory (6): By waiting until the suitor has achieved a high sexual drive state, heightening his arousal by introducing momentary frustration, and then finally rewarding him, the hard-to-get woman can maximize the impact of the rewards she provides.

3. Social Perception Theory: Individuals use information as to another's social standing on one trait as a clue to his standing on related characteristics. For example, individuals may have discovered that very socially desirable dates are harder-to-get than undesirable partners. The two concepts ("hard-to-get" and "socially desirable") might thus become associated. As a consequence, if a girl can successfully simulate being hard-to-get, she may be able to improve others' perception of her desirability.

The first two theoretical explanations of the hard-to-get phenomenon suggest that playing hard-to-get should alter only the suitor's perception of the hard-to-get romantic partner. Social Perception Theory suggests that the hard-to-get individual should impress an even wider constituency. Not only potential suitors, but uninvolved observers as well, should perceive the hard-to-get person as especially socially desirable.

The above rationale leads one to hypothesize that the more romantic interest a stimulus person expresses in a given romantic partner, the less socially desirable that stimulus person will be judged to be by an outside teen-age observer.

ALTERNATIVE HYPOTHESIS

An alternative, and somewhat more complicated, hypothesis also may be proposed. It could be argued that a stimulus person might gain or lose stature by expressing romantic interest in another, depending on how socially desirable the other is.

This hypothesis follows from research by Goffman (4), Kiesler and Baral (5), Dion and Berscheid (2), and Walster and Walster (7), demonstrating that individuals prefer romantic partners of approximately their own level of "social desirability."[2] If teen-agers assume that attractive people are most likely to express romantic interest in attractive others, while the unattractive will only admit to liking the unattractive, the teen-agers might use such

information in evaluating others. Thus, if an ostensibly attractive person expresses great romantic interest in an undesirable partner, the lover should lose stature as a consequence of his liking, while the beloved should gain stature. An ostensibly unattractive person who expresses romantic interest in an attractive partner should gain stature by his liking; the beloved should lose stature.

Thus, one may hypothesize that the attractiveness of a stimulus person, the attractiveness of his partner, and the extent of his romantic interest for the partner, should all be important determinants of how socially desirable the stimulus person and the partner appear to be to an outside observer.

The two experiments reported here were designed to investigate whether or not the knowledge that a person was hard-to-get affected a teen-ager's evaluation of that person. Experiment I varied the attractiveness of the stimulus person, the attractiveness of his partner, and the amount of romantic interest the stimulus person expressed for the partner. In Experiment I, the stimulus person and his partner were always of opposite sexes. Experiment II was similar to Experiment I in every detail, with the exception that the sexual similarity of the stimulus person and his partner was systematically varied.

METHOD—EXPERIMENT I

Subjects and Procedure

Subjects were 144 high school juniors and seniors who belonged to various youth groups in the Rochester, New York area. They were paid $2.00 each for their participation.

To provide a rationale for asking the Ss to rate other students, the experimenter said she was investigating factors which may affect romantic attraction. She especially stressed at this and on several other occasions during the experiment that she was interested in romantic liking and not simply in "friendship." After this introduction, E handed out a detailed description of the project and read it aloud to the Ss:

We'd like you to help us in setting up a study we'll be running in the Fall.

We're interested in the kind of first impression that various teen-agers make on others. We've obtained pictures and some background information about several graduating seniors. We got this information in the following way: We introduced two students who didn't know each other, and had them meet together four times. Then, they filled out a confidential report about their impressions of, and their feelings about each other. As would be expected, some couples really liked each other and others really disliked each other. Also, there were couples in which one person liked the other, but was not liked in return.

Today we'd like to try out some of the materials we plan to use in the Fall. I'll show you pictures of two students, and give you whatever information we have about them, including how they reacted to one another. In some cases we

don't have all the information we'd like as yet, so you'll just have to bear with us.

There are three things we would like you to do. First, go through the booklet and read all the information about both students. Don't answer any questions right away. Instead, think about both of them for a few minutes. Try to imagine what they're both like, how they'd act with one another, and so forth. Then, give us your honest impressions of them. Don't tell us what you think you should think, or what other people would think. Just tell us what you think. Don't hesitate to use the extremes in rating if they seem applicable.

After you've answered a question, you can comment on the question itself, if you wish. If you feel it is unclear, or should be put another way, then make a note on your sheet suggesting how it might be improved.

Subjects were then given a booklet containing the picture and biography of one male and one female student. Half of the time the stimulus person depicted in each photograph was physically attractive, the remainder of the time he was ugly. Beneath each picture was a paragraph describing the school activities of the person depicted. If the person was attractive, the background information implied that he was a very socially desirable individual.[3] For example, the attractive boy's biography said:

Bill is 17 and graduated this June from a New York high school. During the past year he was an active participant in extra-curricular activities at his school. He was a class officer, a member of the football team, one of the editors of the school yearbook, and a member of the band. His hobbies include sports at which he has unusual natural abilities. Bill is also an officer in one of his community's youth groups. He plans to study medicine for his future career.

If the stimulus person depicted in the photograph was physically unattractive, the background information indicated that he was not socially desirable. For example, the ugly boy's biography said:

Jack is 17 and graduated this June from a New York high school. During the past year, he was not an active participant in extra-curricular activities at his school, but he did help to sell school yearbooks and was a member of the band. Outside of school he does some swimming and team sports, although he does not have too much skill at them. Occasionally Jack attends meetings of one of his community's youth groups.

Finally, Ss were told how romantically interested the first stimulus person was in the partner after they had met with each other four times. The stimulus person was said to have liked the other extremely much, not particularly much, or no liking information was provided. If the stimulus person was "extremely romantically interested" in his partner, the following paragraph was added to his biography:

At the conclusion of their four meetings together, Bill was asked to tell us honestly how much liking he felt for Nancy, and how much time he

would be interested in spending with her in the future. He said (1) he liked her extremely much, and (2) that he would enjoy spending a great deal of time with her in the future.

If he was not to be particularly interested in his partner, the last sentence read:

He said (1) he did not particularly like her, and (2) that he would not want to spend time with her in the future.

If the stimulus person's liking for his partner was to be unknown, the sentence read:

We do not have information about whether he likes or dislikes Nancy.

Subjects were never told how much the partner liked the stimulus person.

The variations just described yielded a 2x2x3 design: Attractiveness of the stimulus person, by Attractiveness of the partner by The stimulus person's romantic interest in the partner. Half the Ss assigned to each cell were male and half were female.

Dependent Variables

After considering the photographs and biographies of the stimulus person and his partner for some time, and imagining what it would be like to associate with both teen-agers, Ss were asked to complete a questionnaire composed of the following ten questions: (1) How popular would stimulus person (SP) be with the girls at your school? (2) How popular would SP be with the boys? (3) How much do you think you would like SP? (4) How likely is it that SP is the kind of person you would want to spend much time with? (5) How physically attractive do you think SP is? (6) How much would you guess the partner (P) likes SP? (7) How likely is it that SP is the kind of person who would want to spend much time with you? (8) How physically attractive do you think P is? (9) How popular would you guess P would be with the students at your school? (10) What clues did you use in making these judgments about each member of the pictured couple? How confident do you feel about your judgments?

Scores on questions 1-6 were summed to form an Index of the stimulus person's Social Desirability. Questions 8 and 9 were summed to form an Index of the partner's Social Desirability. (The lower the score on each index, the more socially desirable the stimulus person was judged to be.)

EXPERIMENT II

Subjects and Procedure

Subjects were 128 high school students from the Rochester area.

As previously mentioned, the experimental design of Experiment I was duplicated in Experiment II with the exception that the sexual similarity of the stimulus person and his partner was systematically varied. This necessitated a modification in the experimental procedure. Although E used the same rationale in Experiment II as in Experiment I, she could no longer plausibly claim to be interested in the factors that affect romantic attraction. It was reasoned, however, the Ss would assume that oppo-

TABLE I

EXPERIMENT I: THE EFFECT OF A STIMULUS PERSON'S ROMANTIC LIKING FOR HIS PARTNER AND THE ATTRACTIVENESS OF THE PARTNER, IN DETERMINING Ss' EVALUATIONS

Stimulus Person's Romantic Liking for Partner	Stimulus Person's Attractiveness	Partner's Attractiveness	Perceived Social Desirability of Stimuli[b]	
			Stimulus Person	Partner
Great Interest	Desirable[a]	Desirable	14.50	3.58
	Desirable	Undesirable	14.17	7.33
	Undesirable	Desirable	18.67	3.08
	Undesirable	Undesirable	17.75	7.08
			$\overline{M}=16.27$	$\overline{M}=5.27$
Unknown	Desirable	Desirable	15.83	3.25
	Desirable	Undesirable	15.17	7.58
	Undesirable	Desirable	20.25	3.83
	Undesirable	Undesirable	18.25	7.25
			$\overline{M}=17.38$	$\overline{M}=5.48$
Great Disinterest	Desirable	Desirable	16.00	3.75
	Desirable	Undesirable	13.67	7.67
	Undesirable	Desirable	21.00	3.92
	Undesirable	Undesirable	19.33	7.17
			$\overline{M}=17.50$	$\overline{M}=5.63$

a. N=12 per cell
b. The lower the number, the more desirable the stimuli

site sex relationships were potentially romantic relationships, while same sex relationships were not.

In this experiment, the stimulus person was either said to like his partner extremely much or be disinterested in further interaction with his partner. The condition in which Ss were given no information regarding the stimulus person's reaction to his partner was not included.

The same pictures and biographies described in Experiment I were used in Experiment II, and the stimulus pictures once again varied in attractiveness. Half the time the stimuli were extremely attractive, half of the time extremely unattractive. Appropriate background information was once again provided, and Ss were asked to answer the same questionnaire administered in Experiment I.

The experimental variations in Experiment II, then, yielded a 2x2x2 design: Sexual similarity of SP and P by Attractiveness of SP by Attractiveness of P by SP's Romantic Interest in P.

RESULTS

Manipulation Check

The social desirability of the stimuli were successfully manipulated. In both experiments, the attractive stimulus person was judged to be more socially desirable than the unattractive person (in Experiment I, F=63.22, in Experiment II, F = 25.56).

The attractive partner was also judged to be more socially desirable than was the unattractive partner (in Experiment I, F = 190.68, in Experiment II, F= 195.05).

Experimental Results

With respect to our hypothesis (that the more romantic interest a person expresses in another, the less socially desirable that person will appear to an outside observer), the data were clear. The results are diametrically opposed to those predicted. From Tables 1 and 2 it is evident that the more interested the stimulus person admits he is in the partner he met a short time before, the more socially desirable teen-agers assume the stimulus person must be. In Experiment I, this linear trend was not quite significant (F = 3.4, 5; p = .07). In Experiment II, however, similar results were secured and this main effect was statistically significant (F = 8.11). The more the stimulus person liked his partner, the more socially desirable teen-agers perceived the stimulus person to be.

The stimulus person's Social Desirability Index was constructed by summing Ss' answers to six questions. Let us examine each of the six questions.

In Experiment I we find that when the stimulus person is romantically interested in his partner he is evaluated more highly on all six questions than when he is disinterested in his partner. On only one question, however, does the difference reach statistical significance. (The more romantic interest SP expressed in P, the more Ss assumed that the stimulus person reciprocated his liking F = 8.70.) In Experiment II, the stimulus person who liked his partner was again rated higher on all six items making up the Social Desirability Index. On only three of these items, however, were there statistically significant main effects. The more the stimulus person liked his partner the more Ss liked the stimulus person

TABLE 2

EXPERIMENT II: THE EFFECT OF THE STIM-
ULUS PERSON'S LIKING FOR HIS PARTNER ON
Ss' EVALUATIONS

Stimulus Person's Liking for His Partner	Sex of the Stimulus Person and His Partner	Perceived Social Desirability of Stimuli[b]	
		Stimulus Person	His Partner
Great Liking[a]	Same Sex	17.07	5.72
Great Liking	Opposite Sex	18.04	6.41
Great Disinterest	Same Sex	19.16	6.44
Great Disinterest	Opposite Sex	19.41	6.57

a. N=32 per cell
b. The lower the number, the more desirable the stimuli.

(F = 9.82), the more time Ss wanted to spend with him (F = 6.10), and the more Ss assumed the partner must have liked him (F = 25.43).

Alternative Hypothesis

With respect to the alternative hypothesis (that whether or not a person gains or loses stature by expressing romantic interest in another depends on the social desirability of the object of his affection) the data are again clear. There is no support for the notion that the attractiveness of the stimulus person, and the degree of the attractiveness of his partner, and the degree of liking SP expresses for P will interact in determining how socially desirable the stimuli are judged to be. The alternative hypothesis predicted that unattractive stimuli would gain stature if they liked or were liked by attractive individuals, and attractive individuals would lose stature if they liked or were liked by ugly individuals. These predicted 3-way interactions were all nonsignificant. First, consider Ss' ratings of the stimulus person's social desirability: In Experiment I, the predicted 3-way interaction equalled .47; in Experiment II, F = .00. When we consider the Ss' ratings of the partner, the results are the same: In Experiment I, the predicted 3-way interaction equalled .22; in Experiment II, F = .14.

The complete rejection of this hypothesis is somewhat surprising. Had the hypothesis been supported, the results would have been consistent with the findings of Kiesler and Baral (5), Dion and Berscheid (2), and Walster and Walster (7). In addition, the results would have been consistent with the common sense observation that individuals assume that they lose stature by liking or being liked by the "wrong" individuals. In informal interviews conducted with several of the high school girls, many confessed that it is extremely embarrassing to be asked out, in public, by socially undesirable boys. Part of the embarrassment probably arises from the fact that when an unacceptable person asks one out, one is faced with the problem of publicly rejecting the undesirable suitor in a tactful way. However, the reason most commonly cited by the teen-agers for being embarrassed when asked out by a "creep," was that "my friends might think that I'd actually go out with some-

one like that!" The girls assumed they would lose status if they liked or were liked by others less desirable than themselves. The data collected in the present two experiments suggest that their fears may be groundless.

In sum, the present data indicate that people simply like people who like people. There is no evidence for the hypothesized effectiveness of a hard-to-get strategy. Both hard-to-get hypotheses failed to receive even a suggestion of support.

FOOTNOTES

1. This research was financed in part by National Institute of Mental Health Grants 16661 and 16729 and in part by the Office of the Dean of Students, University of Minnesota. We would like to thank Elaine Rosenwasser for running this experiment.

2. "Social Desirability" was defined by Walster and Walster (7) as "The sum of an individual's social assets, weighted by importance and salience for others." Social assets such as physical attractiveness, popularity, personableness, and material resources were presumed to be important factors in determining one's social desirability level.

3. An experiment was run with Rochester high school seniors to insure that the photographs and biographies of the "socially desirable" stimuli were perceived as more desirable than were the photographs and biographies of the less desirable stimuli.

4. In Experiment I, df = 1 and 96. In Experiment II, df = 1 and 112.

REFERENCES

1. Crowne, D. P.; Marlowe, D., The Approval Motive, Wiley, New York, 1964.

2. Dion, Karen; Berscheid, Ellen, "Physical Attractiveness and Dating Choice: A Test of the Matching Hypothesis," available in mimeograph, 1969.

3. Festinger, L., A Theory of Cognitive Dissonance, Row, Peterson, Evanston, Illinois, 1957.

4. Goffman, E., "On Cooling the Mark Out: Some Aspects of Adaptation to Failure," Psychiatry, 15:451-463, 1952.

5. Kiesler, Sara B.; Baral, Roberta L., "The Search for a Romantic Partner: The Effect of Self Esteem and Physical Attractiveness on Romantic Behavior," available in mimeograph, 1967.

6. Kimball, G. A., Hilgard and Marquis' Conditioning and Learning, Appleton-Century-Crofts, New York, 1961.

7. Walster, Elaine; Walster, G. W., "The Matching Hypothesis," available in mimeograph, 1969.

8. Xenophon, E. C. Marchant (trans.) Memorabilia, III, xi, Heinemann, London, 1923.

ESTIMATED EFFECTS OF FOUR FACTORS

ON ACADEMIC PERFORMANCE

BEFORE AND AFTER TRANSFER

SAM C. WEBB
Georgia Institute of Technology

ABSTRACT

This study proposes an analytical design for assessing the contribution of four factors to academic per-
formance of students transferring from one college to another. Average grades " before " transfer and for var-
ious periods " after " transfer were computed for the transferring students and for a matched sample of native
students. Using these averages, the contribution of four factors—differential grading standards, preparation for
advanced work, academic potential, and coping ability—to the difference between the average " before " transfer
and the average for the first quarter " after " transfer for the transferring students were estimated. In the il-
lustrative data, the largest contributor to an observed decrement of 1. 21 letter grades was differential grading
standards. The factor seemingly contributing least to the decrement was preparation for advanced work.

WHILE VARIOUS procedures (10, 12, 18) are
available for studying the academic performance of
college transfer students, the approach usually fol-
lowed emphasizes the comparison of academic per-
formance before and after transfer and/ or the com-
parison of academic performance of native and transfer
students. A number of measures such as honors
at graduation (7), attrition rates (11, 13), gradua-
tion rates (9, 14), time of graduation (15), and pro-
portion on probation (17, 18), have been employed.
Most studies following the before and after approach
have considered primarily earned grade point aver-
ages (GPA'S) as a criterion measure.

Some investigations (1, 2, 4, 17, 18) have consid-
ered students transferring from one 4-year college to
another; but the vast majority of studies have been
based on junior college students transferring to se-
nior institutions. Transfers from other kinds of in-
stitutions appear to perform better in senior college
than do transfers from junior colleges (5); but the
data are too scant to make firm generalizations.

While there are variations in results as reported
from study to study, several strikingly consistent

trends in the performance of junior college transfers
at senior colleges are evident. For example, there
is decline in average grades immediately after trans-
fer ((5) forty-four of forty-six studies; (10), forty
of forty-one studies). Average grades improve in
subsequent quarters ((5) thirty-four of thirty-eight
studies; (10) thirty-nine of forty - one studies).
But the 2-year cumulative average does not exceed
the before transfer average ((10) forty of forty-one
studies). Native students tend to earn higher grades
than transfers ((5) twenty-two of twenty-three
studies), while transfers have higher attrition rates
(13) and graduate later and in smaller proportions
((5) nineteen of twenty-one studies).

With such reliability across studies, there can
be little doubt that these generalizations faithfully
describe the academic performance of junior college
transfers as measured in terms of average grades.
However, since results are usually obtained by av-
eraging grades of all transfer students at a given
college or by averaging grades of all students from
specified types of colleges, the results are descrip-
tive of students from no particular college. Further,
by virtue of their poor design, reported studies

seldom assist in assessing the role of various factors which may be influential in producing the results described.

For example, several studies show that ostensibly significant differences in averages favoring native students are substantially reduced or disappear when differences in academic potential are taken into account (8, 9, 11). Further, Knoell and Medsker (10:96-98) have illustrated the influence of institutional characteristics, non-intellective characteristics of students, and general cultural factors on after transfer performance by demonstrating significant relationships of type of 4-year college, sex, and state differences to after transfer performance.

These and other findings suggest that after transfer performance may be jointly determined by various environmental and student characteristics acting together, so that by virtue of the singular combination of these factors for any given pair of colleges—one from which students transfer and one to which students transfer—the combination of factors and their relative degrees of importance in determining after transfer performance for the pair may well be unique. Suggestive of this possibility is a finding by Willingham (18) showing that optimal correctional weights for predicting grades at Georgia Tech for transfer students classified into sixteen homogenous groups ranged from -1. 1 to 0. Also suggestive is a report on nineteen Florida junior colleges indicating that average grades after transfer ranged from an increase of 0. 02 of a letter grade to a decrease of 0. 74 of a letter grade (1).

In summary, findings suggest that while a variety of factors may relate to the after transfer performance of college students, presently available studies offer little assistance in identifying them or in assessing their relative influence on such performance. Further by virtue of the variety of—and even unique combinations of—such factors that may be found in pairs of colleges, studies based on students transferring from one single school to another single school may be helpful in understanding the dynamics underlying the academic performance of transfer students.

PURPOSE

The present study attempts to assess the contribution of four factors as they affect the comparative performance before and after transfer for a group of students who transferred from one junior college to one 4-year college. The factors considered are grading standards, preparation for advanced work, academic potential, and coping ability. It was expected that the results would provide a better understanding of the dynamics of transfer from the one school to the other and thus make possible better guidance for students who might consider transferring. In addition, results might facilitate the transfer process for students who do transfer.

METHOD

In reporting this study the school from which students transferred is called the feeder school while the school to which they transfer is called the host school. Students (N = 130) (subsequently called "transfers")

who transferred from a feeder school with a curriculum closely resembling that of the host school were selected for study. They all transferred during the period from fall 1961 through fall 1965. Records were studied through spring of 1966.

A comparison group of students at the host school (hereafter called " natives ") was selected from the 1961 freshman class to match the transfer students on a person-for-person basis primarily in respect to academic potential, and secondarily in respect to quarters of enrollment in school. Matching in respect to academic potential was made on the basis of the verbal (SAT-V) and mathematical (SAT-M) scores of the Scholastic Aptitude Test, with emphasis being given to exact matching on the latter score. The performance record of each selected native student was then divided into " before "and" after " transfer segments so as to be equivalent formally to the segments of the record of his transfer counterpart. This division was accomplished by dividing the record between two quarters so that the total credit hours for the " before" transfer segment would closely approximate the transferred credit for the transfer student counterpart. Further the " after " transfer segment was truncated so that the number of quarters after transfer did not exceed the number of quarters attended after transfer by the transfer student counterpart.

Summary statistics for the matching variables and for several other variables descriptive of the groups are shown in Table 1. The predicted average (PA) was obtained by use of a linear equation for predicting first year averages at the host school from SAT scores. It was derived from data for the entire entering freshman class. Predicted and earned averages were computed on the basis of a scale in which A = 40, B = 30, C = 20, D = 10 and F = 0.

For both groups of students average grades for the before transfer segments of their records for verbal subjects (English and social science courses), for quantitative subjects (chemistry, mathematics, and physics), and for all work taken (excluding physical education, band, etc.) were computed. These average values and their differences obtained by subtracting native from transfer averages are shown in the top half of Table 2.

Also average grades for various segments of work after transfer were computed for both groups. These included all courses taken in the first quarter, verbal subjects taken in the first three quarters, quantitative subjects taken in the first three quarters, all work taken in the first three quarters, all work taken in quarters 4 through 6, all work taken in quarters 7 through 9, and the cumulative average for all work taken. These averages and their differences are shown in the lower portion of Table 2.

Finally, estimates of the contribution of the four factors of concern to the comparative before and after transfer performance for the transfer students were computed by methods subsequently described.

RESULTS

Comparability of Groups

In selecting the matching native students,

TABLE 1

DESCRIPTIVE DATA FOR MATCHING TRANSFER AND NATIVE SAMPLES

Measures	Transfers			Natives			Difference
	N	M	SD	N	M	SD	
SAT-V	130	425	85	130	442	75	-17
SAT-M	130	514	77	130	519	69	- 5
High School Average	40	29.8	6.1	130	30.3	5.5	- 0.5
Predicted Average	130	19.0	3.0	130	19.3	3.1	- 0.3
Hours Credit Transferred	130	77.8	24.4	130	73.3	31.1	4.5
Total Quarters Enrolled	130	11.2	3.47	109	10.2	3.16	1.0*
Quarters Enrolled before Transfer	130	5.7	1.58	117	5.2	1.57	0.5*
Enrollment by Curriculum [a]							
Engineering	72			54			18*
Industrial Management	18			34			-16
Science and Architecture	10			12			- 2
Status at End of Study [a]							
Graduated or Still in School	63			76			-13*
Withdrew or Dropped	37			24			

* Significant 5 percent Level of Confidence.

[a] Reported in percents.

emphasis was given to selecting a group who would, on a person-for-person basis, be as near like the transfer students in respect to academic potential as possible. Since the differences between groups in respect to SAT-V, SAT-M, high school average, and predicted average were not significant (Table 1), this goal seems to have been achieved. Similarly there was no significant difference in respect to hours of credit transferred. It must be noticed, however, that the groups do differ significantly at a 5 percent level of confidence in respect to total quarters enrolled before transfer, choice of engineering as a curriculum, and proportion still in school or

TABLE 2

SELECTED GRADE POINT AVERAGES FOR TRANSFER AND NATIVE SAMPLES

Measures	Transfers			Natives			Difference
Before Transfer	N	M	SD	N	M	SD	
Verbal Average	129	25.3	5.8	129	18.7	6.2	7.0*
Quantitative Average	130	27.6	6.3	128	15.6	8.5	12.0*
Total Average	130	27.9	4.8	129	19.4	5.9	8.5*
After Transfer							
First Quarter Average	127	15.8	7.9	109	22.7	6.3	-6.9*
Verbal Average (Quarters 1-3)	99	20.3	7.3	103	22.4	6.6	-2.1*
Quantitative Average (Quarters 1-3)	113	13.2	9.5	98	20.5	9.6	-7.3*
Total Average (Quarters 1-3)	122	17.2	6.3	109	22.6	5.7	-5.4*
Total Average (Quarters 4-6)	79	21.9	5.0	65	24.8	5.6	-2.9*
Total Average (Quarters 7-9)	49	24.2	5.2	34	27.0	4.5	-2.8*
Final Average	127	18.5	8.2	109	23.2	5.5	-4.7*

* Significant 5 percent Level of Confidence.

graduated at the time of termination of the study. However, it was not possible to make a closer match on these variables without creating significant differences on the measures of academic potential. Also it is relevant to note that mean SAT scores for the matching sample of native students were 77 points lower on SAT-V (442 versus 519) and 73 points lower on SAT-M (519 versus 592) than average scores for the entire freshman class from which the sample was selected.

Comparative Performance of Transfer and Native Students

While the details will not be cited, the data in Table 2 show that even when the two groups were matched in respect to academic potential, the results follow the pattern typically found in transfer studies. Transfer students made higher averages than natives before transfer. After transfer there was a decrement in the performance level of the transfer group. Though average grades after transfer for the transfers showed gradual improvement after the first quarter, and though differences between averages for natives and transfers became smaller, averages for the transfers never equalled or exceeded the averages of the natives in comparable segments of the period studied; neither did they reach the level attained before transfer. In contrast, natives showed no decrement in performance " after transfer "; and performance

TABLE 3

SUMMARY OF ESTIMATED CONTRIBUTION OF FOUR VARIABLES TO DECLINE IN AVERAGE GRADES AFTER TRANSFER

Source	Method of Estimation	Amount
Grading Standards	Before Transfer, (Native)- Before Transfer, (Transfers)	-8.5*
Coping Ability Conservative:	After Transfer, First three Quarters- After Transfer, First Quarter (Transfers)	-1.4*
Dashing:	After Transfer, Quarters 4 through 6- After Transfer, First Quarter (Transfers)	-6.1*
Preparation Plus Coping Ability		
Conservative:	Before Transfer, (Native)- After Transfer, First Quarter (Transfer)	-3.6*
Dashing:	After Transfer, First Quarter (Native)- After Transfer, First Quarter (Transfers)	-6.9*
Preparation		
Conservative:	(Preparation and Coping)- Coping	-2.2[a]
Dashing:	(Preparation and Coping)- Coping	-0.8[a]

(Continued)

(TABLE 3 Continued from Column 1)

Source	Method of Estimation	Amount
Academic Potential	Average Grades, Before Transfer (Native)- Average Grades (Total Host)	-3.2[a]
	Estimated Total (Dashing) Sum of Above	-18.6
	Estimated Total (Conservative) Sum of Above	-15.3
	"Actual" Total After Transfer, First Quarter (Transfers)- Before Transfer (Transfers)	12.1

* Significant at 5 per cent Level of Confidence.
[a] Significance Not Tested.

levels increased with successive periods of enrollment.

Contribution of Four Factors to Grade Decrements After Transfer

Table 2 indicates that the transfer group showed a decrement of 12.1 (1.21 letter grades) from the average before transfer to the first quarter after transfer. The following paragraphs are devoted to estimating from the data in Table 2 what portion of this decrement can be associated with each of four factors: grading standards, preparation for advanced work, academic potential, and coping ability. The methods of estimation and the resulting estimates are shown in Table 3.

Grading Standards. The grading standard refers to the judgmental scale instructors use in evaluating (assigning grades to) the performance of their students. Operationally it may be defined as the average grade assigned to the work of students whose academic potential is at a specified level. Since the academic potential of the transfer and native groups is the same, an estimate of the decrement attributable to the difference in grading standards for the two schools is given by the difference between the before transfer averages at the two schools. Thus, a difference of 8.5 points can be attributed to differences in grading standards, indicating, of course, that the grading standard at the host school is that much lower or harder than the grading standard at the feeder school.

Coping Ability. Coping ability refers to the kinds of behaviors required of students for dealing adequately with a given environment. If the presence or absence of such abilities affect academic performance, grades should decline when one moves from one environment to another, especially if the environments are substantially different; and grades should improve as one learns more adequately to cope with a given environment. An estimate of the effect of improved coping ability can be made by subtracting the average for the first quarter after transfer for the transfer students from the average in some subsequent period. Following this procedure,

a conservative estimate of the effects of improved coping ability on grades for the transfer students can be obtained by subtracting average grades for the first quarter after transfer from average grades for the first three quarters after transfer. A more "dashing" estimate can be made by subtracting average grades for the first quarter after transfer from average grades for quarters 4 through 6. These yield values of 1.4 and 6.1 respectively. By assuming that transfer from one environment to another produces a reverse effect, one estimates the decrement attributable to coping effects equals -1.4 (conservative), and -6.1 (dashing).

Preparation for Advanced Work. Preparation for advanced work refers to the adequacy of the student's knowledge and understanding as required for satisfactory comprehension of materials in courses subsequently to be taken. An estimate of the decrement immediately following transfer attributable to this factor can be obtained only indirectly. A "dashing" estimate of the decrement after transfer attributable to the combination of inadequate preparation and inadequate coping ability can be made by subtracting the first quarter average after transfer for the transfers from the first quarter average after transfer for the natives. According to this procedure there is a decrement of 6.9 attributable to the combination of these two factors. This estimate may be somewhat attenuated by the fact that while for the transfer students, there was a reduction in grades following transfer as a function of reduced coping ability, an increase in grades following "transfer" might be expected for the natives as a function of improved coping ability. In fact, an increment of 3.3 occurred for this group.

An estimate of the combined effects of coping ability and preparation which obviates this complication and which may thus be considered as more conservative may be obtained by subtracting the first quarter after transfer from the total average before transfer for the natives. This procedure yields a decrement of -3.6.

With these estimates of combined effects, available estimates of the decrement associated only with preparation effects can be obtained by subtracting the estimate for coping ability effects from these values. This procedure yields dashing and conservative estimates of decrements for the first quarter after transfer of 0.8 and 2.2 respectively.

Academic Potential. Academic potential refers to the intellectual abilities and aptitudes usually found associated with college grades. Since the two groups used in the study were matched in respect to academic potential, differences in grades between groups and differences in grades before and after transfer for the respective groups, superficially at least, would not appear to be associated with differences in academic potential.

However, it is relevant to recall that, except in rare instances, grading standards are usually formulated in such a way as to keep the distribution of grades about the same even though the levels of academic abilities may change, (3:9-19, 16). At the host school for example, while mean SAT-V and SAT-M scores for entering freshman classes increased 74

and 88 points respectively from 1957 through 1965, earned average first quarter grades increased only 0.03 of a letter grade.

While procedures are available for estimating the changes in standards that accompany changes in ability levels (6, 16), they are not altogether applicable to the data that have been presented. However, it is relevant to note that the median SAT-V and SAT-M scores for the feeder school fell at the nineteenth and seventeenth percentile ranks for the entering 1961 freshman class at the host school. Thus students who form the ground (6) in terms of which evaluations are made at the feeder and host institutions are quite different in respect to academic potential. Consequently it seems reasonable that the decrement in grades between the first quarter after transfer and the average before transfer for the transfer students may be partly a function of these differing backgrounds. An estimate of the decrement attributable to these different perceptual grounds can be obtained by subtracting the average grade before transfer for the native students from the average grades for the entire 1961 freshman class of the host school. Unfortunately the exact data required for this computation are not available. But an approximate estimate can be made by using average first quarter grades for the freshman year. For the entire freshman class and for the matching natives this value is 19.5 and 16.3 respectively. These values then yield an estimated decrement of 3.2.

Combining the several estimates for the four variables, the conservative estimated total decrement equals 15.3 and the "dashing" estimated total decrement equals 18.6. These total estimates are respectively 3.2 and 6.5 higher than the actual decrement of 12.1; they are hence attenuated to that extent.

SUMMARY AND DISCUSSION

The foregoing presentation has in essence described and illustrated the use of a proposed design for estimating the contribution of four factors which are contributions to the difference in academic performance as measured by average grades for students who transfer from one school to another. While the data reported here deal with a decrement in performance following transfer, the design and estimating procedures are believed to be general in application in the sense that they can be applied with equal validity to the analysis of gains as well as decrements in achievement following transfer. It may well be, however, that the estimating procedures are sufficiently imprecise that measurable increments or decrements can be identified only when the total gain or decrement is fairly large—as in the particular situation reported here. At any rate, the design permits the possibility that the pattern of changes associated with the four factors may be strikingly different as among various school pairs.

As for the actual estimates obtained, cognizance must be taken of the fact that the sum of decrements obtained by both the conservative and dashing estimating procedures exceed the actual decrement. These overestimates may arise from the fact that the estimating procedures are in themselves faulty and/or from the present inability to identify and correctly

assess the interactions of the four variables considered. For example, it will be noticed that the conservative sum of decrements for the four factors exceeds the observed total decrements by 3.2 an amount which equals precisely the estimated decrement for academic potential. This observation is suggestive of the possibility that this measure is subsumed in the estimated decrement attributed to differences in grading standards.

Another point of potential error that may be identified is that of estimating the combined effects of preparation plus coping ability. In addition to the two estimates already noted, estimates based on later segments of the performance record can be made. For example, estimates for the first three quarters, for quarters 4 through 6, and for quarters 7 through 9 are -5.4, -2.9, and -2.8 respectively. The observed reduction in difference for the successive time periods is consistent with the expectation that as the transfers experience more and more of the environment and of the same level of instruction as that experienced by the natives, differences attributable to these factors would tend to disappear.

The fact that the observed decline in decrement may not be entirely a function of improved preparation and coping ability for the transfer students is suggested by the increasingly smaller numbers of students as a function of the elimination of poorly achieving students in the successively later periods of enrollment (Table 2).

It is possible to estimate the effect of this phenomenon on some of the differences attributable to the combined effects of preparation and coping ability already noted by taking into account for the comparison periods, the average grades made in each of these periods by the forty-nine transfer students who persisted through seven to nine quarters. According to the results shown in Table 4, it would appear that approximately half the difference for each period considered can be attributed to this phenomenon.

Other complexities and complications which have an effect on the accuracy and interpretation of the several estimates can also be noted, but not easily assessed. For example, the validity of the interpretations given the estimates are predicated upon the assumption that the effects of the several factors considered remained more or less constant—i.e., operated with more or less the same strength—throughout the period studied. There is evidence to suggest such an assumption is not altogether valid. For example, there are some reasons to believe that standards at the host institution may have been easier in the last 2 than in the first 2 years, and at the same time over the period studied the general level of the standard may have gone up or become harder. While it is difficult to assess precisely the effects of these changes on the reported results, nevertheless it is clear that a difference in standard between the first 2 and last 2 years at the host school could partially account for the relatively large difference between the before transfer and the first quarter after transfer average for the native group.

Similarly difficulty results from the fact that in the period covered by the study, students in both groups dropped out, but not in pairs. For example, native students withdrew or were dropped in greater numbers before transfer, while transfer students withdrew or were dropped in greater numbers after transfer. These trends would tend to increase the difference between the before transfer and first quarter after transfer average for the natives and increase the differences between averages in the after transfer segment for transfers.

A final complicating factor to be noted relates to divisional and departmental differences as they affect earned GPA's. From Table 2, for example, it is evident that the grading standard at the host institution is harder than that of the feeder institution by half a letter grade and one and a half letter grades for verbal and quantitative subjects respectively. This kind of difference has been noted by other investigators (15). These complexities serve only to remind one of the difficulty of finding in the educational setting data that may be clearly and unequivocably interpreted. For the present study at least, the estimated values are of sufficient magnitude to demonstrate an effect of the factors considered on the comparative performance of the native and transfer students.

The possible ambiguous interpretation of some of the data, however, suggests the need for further efforts directed toward the development of more precise analytical designs that can increase precision of measurement and identify possible interactive effects.

However, the present estimates, even though approximate, seem potentially useful in understanding more adequately the contribution of factors underlying the expected changes in student performance following transfer from one school to another. Were such information available, for instance, to advisers of a given feeder school relative to the several host schools to which its students transfer, they would be able to counsel with students considering transfer in a far more effective way than is now possible. Also, on the basis of such information, school administrators should be able to work out more effective procedures for minimizing difficulties in the transfer process.

TABLE 4

ESTIMATED CONTRIBUTIONS OF DROPOUT STUDENTS AND OF PREPARATION AND COPING ABILITY TO DIFFERENCES BETWEEN SELECTED AVERAGES FOR TRANSFER AND NATIVE STUDENTS

Source of Difference	Time Period		
	First Quarter	First Three Quarters	Quarters 4-6
Dropout of Poorly Achieving Students	3.2	3.2	1.5
Preparation and Coping Effects	3.7	2.2	1.4
Total	6.9	5.4	2.9

FOOTNOTE

1. Portions of this paper were presented as part of a program on " The Transfer Student's Academic Success " at the Annual Meeting of the Association for Measurement and Evaluation in Guidance, Dallas, Texas, March 22, 1967.

REFERENCES

1. " An Abstract of Findings: The Academic Performance of Florida Junior College Transfer Students in Florida Degree Granting Institutions, Fall Term, 1959," (undated).

2. Ahmann, J.S., " Academic Attrition of Engineering Transfer Students," The Journal of Experimental Education, 24: 141-146, 1955.

3. College Student Profiles, 1966-67 Edition, American College Testing Program, Iowa City, Iowa, 1966.

4. Godfrey, R., " An Institutional Study Related to the Admission of Transfer Students," presented at the Annual Meeting of the Southern Association of Collegiate Registrars and Admission Officers, Memphis, Tennessee, 1963.

5. Hills, J.R., " Transfer Shock: The Academic Performance of the Junior College Transfer," The Journal of Experimental Education, 33:201-215, 1965.

6. Hills, J.R.; Gladney, M.B., " The Study of Factors Influencing College Grading Standards," Testing and Guidance Research Bulletin 2-66, Board of Regents, University System of Georgia, Atlanta, 1966.

7. Holmes, C.H., " The Transfer Student in the College of Liberal Arts," Junior College Journal, 31: 456-461, 1961.

8. Jones, F.M., " A Controlled Comparison of the Academic Performance of Native and Transfer Students at the University of Georgia," unpublished doctoral dissertation, University of Georgia, Athens, 1966.

9. Klitzke, L.L., " Academic Records of Transfers in Teacher Training," Junior College Journal, 31: 255-257, 1961.

10. Knoell, D.M.; Medsker, L.L., " Factors Affecting Performance of Transfer Students from Two-to-Four-Year Colleges: With Implications for Coordination and Articulation," Cooperative Research Project No. 1133, University of California, Berkeley, 1964.

11. Lindsay, C.A.; Marks, Edmond; Hamel, L.S., "Native and Transfer Baccalaureate Students," Journal of College Student Personnel,pp. 5-13, 1966.

12. Linn, Robert L., " Grade Adjustments for Prediction of Academic Performance: A Review," College Entrance Examination Board Research and Development Reports RDR-64-5, No. 18, Educational Testing Service, Princeton, New Jersey, 1965.

13. Medsker, L.L., The Junior College: Progress and Prospect, McGraw Hill Book Co., New York, 1960.

14. Ruch, G.M.; Baker, D.G.; Ryce, E.,"A Comparison of the Scholarship Records of Junior College Transfers and Native Students of the University of California," California Quarterly of Secondary Education, 4:201-213, 1928.

15. Russell, J.W., " The Analysis of the Academic Performance of Transfer and Native Students and Their Major Fields in the College of Arts and Sciences at the University of Georgia," unpublished doctoral dissertation, School of Education, University of Georgia, Athens, 1963.

16. Webb, S.C., " Measured Changes in College Grading Standards," College Board Review, 39: 27-30, 1959.

17. Webb, S.C., " An Evaluation of the Potential of Transfer Applicants," Evaluation Studies Research Memorandum 65-2, Georgia Institute of Technology, Atlanta, 1965.

18. Willingham, W.W., " Prediction of Academic Success of Transfer Students," Evaluation Studies Research Memorandum 61-66, Georgia Institute of Technology, Atlanta, 1961.

EFFECTIVENESS OF INSTRUMENTAL AND TRADITIONAL METHODS OF COLLEGE READING INSTRUCTION

RICHARD P. WHITEHILL and SUE J. RUBIN
The University of Wisconsin

ABSTRACT

Introverted and extraverted Ss were assigned to traditional and instrumental reading groups. The performance of Ss assigned to the instrumental treatment condition was superior to those in the traditional condition. No significant main effect was found for the introversion-extraversion dimension, although the performance of instrumentally trained extroverts was superior to that of other groups.

WHITEHILL AND Jipson (4) demonstrated the role of the introversion-extraversion (I-E) variable in college developmental reading program performance. Using instrumental and traditional reading program formats, Ss were sorted on the basis of I-E scores and their performance increments were compared. There were two prime findings. First, the instrumental program appeared on the whole more effective than the traditional program as measured by gain scores accrued across all groups of Ss. Second, extraverts (E) were affected to a greater extent than introverts (I) by the program variable. That is, there was little overall difference within the introvert group between traditional and instrumental programs, while there were significant differences within the extravert group on the program variable, with the extraverts accruing better gains in the instrumentally based program. The purpose of the present paper is to replicate the Whitehill-Jipson study with particular attention to the overall program variable.

METHOD

Subjects

The Ss for this study were forty graduate and undergraduate students at The University of Wisconsin during the 1969 spring semester. Subject selection took place after administration of the Eysenck Personality Inventory (EPI) during the first session of a free, voluntary developmental reading course. The scores on the EPI were divided into thirds; Ss ranking in the upper 33 percent of the cases composed the E groups; those in the lower 33 percent the I groups, and those in the middle 33 percent the M or "normal" groups. The EPI was given to all students who registered for the course. A test of current reading ability, the Cooperative English Test, Form A, was also administered to all students entering the program. Ss were matched according to sex, age, year in school, college curriculum within the university, scores on the EPI, and comprehension percentile score on the Cooperative English Test. They were then randomly assigned to an experimental or control condition; six E's, seven I's, and seven M's were taught by traditional methods; six E's, seven I's, and seven M's were taught by experimental methods.

Apparatus

The experimental group used automatic reinforcing clocks which were equipped with a light that went out when a given criterion speed on a 500-word passage was not met.

Traditional Ss alternated between the use of a Science Research Associates (SRA) Reading Accelerator and timed readings on stop watches. The accelerator is a device with a bar which moves down the page, covering material at a preselected rate. The student is forced to increase his speed to read the material before it disappears from sight.

The reading material for the experimental Ss consisted of paperback books divided into 500-word passages. The experimental Ss were in the traditional class for their initial three lessons, after which they were transferred to the operant method and Hiroshima by John Hersey. Upon completion of that book, they read The Bridge at Andau by James Michener. Thereafter, Ss were free to choose any paperback they wished at a level of difficulty equivalent to these books.

Traditional group Ss used the texts Increasing Reading Efficiency (3) and Maintaining Reading Efficiency (4). These workbooks contain reading selections on various topics. Each selection is followed by ten short answer or objective comprehension questions. The difficulty level of the experimental and traditional material was judged to be equivalent.

Procedure

All students read selection No. 1 in Efficient Reading (1) during their orientation session; this measure provided a common pretest of basal reading speed and comprehension. The traditional group attended two 50-minute classes per week for as many weeks as desired. They used the traditional texts, completing twenty selections in Increasing Reading Efficiency and a varied number in Maintaining Reading Efficiency until a speed of 1,000 words per minute (WPM), timed on a stop watch, was easily attained. At that point, Ss were permitted to read a paperback or their own material using the same methods. The ten comprehension questions for each of the text selections were multiple-choice, fill-in-the-blank type; they were answered after each selection and scored with 10 points for each correct answer. A minimal score of 60 percent was considered acceptable in terms of comprehension.

The experimental group also attended two 50-minute sessions per week. During each session, S read as many 500-word passages as time permitted. He was provided with a chart of time clock settings giving WPM rates. If S's reading speed for a particular passage was slower than the given criterion, the light went out at the end of the time period before S had completed the passage. If S did not complete the passage, he was instructed to read subsequent passages at the same rate until criterion speed was met. Each time S successfully read at criterion speed, he stopped the clock, recorded his success, and raised the criterion 50 WPM.

Five comprehension questions accompanied every four 500-word passages. Subjects were instructed to answer each question upon completion of the corresponding textual sections. The questions were parallel to the traditional group questions in that they dealt with factual recall, inference, and generalization from the reading. The experimenter graded these questions between sessions and extra instruction was given to Ss who could not maintain 60 percent comprehension.

RESULTS

All Ss completed at least six sessions. Individual records were kept of each S's beginning WPM rate and final WPM rate on a session by session basis.

Analysis of pretreatment WPM scores show no significant differences between any of the S groups. Analysis of personal data with regard to age and year in school of Ss also did not yield significant results.

Table 1 shows the mean percentage gain in WPM scores between the first and sixth sessions for each group and treatment. As a whole, the experimental treatment groups made more progress, attaining a 233 percent mean gain in WPM as compared to the 163 percent mean gain for the traditional treatment.

TABLE 1

MEAN PERCENTAGE GAIN IN WPM BETWEEN FIRST AND SIXTH SESSIONS

Treatment	E	M	I	Mean
Traditional	159.2	192.6	137.7	163.3
Experimental	272.2	210.7	221.1	232.8
Mean	215.7	201.6	179.4	

An examination of the performance of I's, E's, and middle Ss in each treatment reveals a similar trend. Each group in the experimental treatment demonstrated a higher mean percentage WPM than the corresponding group in the traditional treatment. The difference is greatest for the E's, and least for the M's. There is 113 percent difference in mean gain between E-ex (extravert-experimental) and E-t (extravert-traditional), 83.4 percent difference between I-ex (introvert-experimental) and I-t (introvert-traditional), and an 18.1 percent difference between M-ex (middle-experimental) and M-t (middle-traditional). Combining treatments, the E's made the most mean percentage WPM gain (215.7%), the M's next (201.6%), and the I's least (179.4%).

A two-way analysis of variance, using the Scheffé approximation and assuming a fixed effects model, was computed using these data. The Scheffé approximation was used because of the unequal numbers in each group. Table 2 illustrates the analysis of variance. There is a significant difference between experimental and traditional treatments as far as mean percentage WPM gain is concerned. An F ratio of 4.02 was obtained, and this is significant at the .05 level. However, no significant difference was found in the I-E variable. Also, the E-ex, I-ex, and M-ex groups did not obtain WPM gains that were significantly different from each other; this relationship also holds for the E-t, I-t, and M-t groups. Furthermore, interaction effects were not significant. A Scheffé multiple comparison between E-ex and I-ex groups yields a corrected F of 23.03; p < .01.

TABLE 2

SUMMARY OF ANALYSIS OF VARIANCE

Source	SS	df	MS	F
Program	5.09	1	5.09	4.02*
I-E	8.86	2	4.43	0.35
Interaction	1.56	2	7.81	0.62
Error Within Cells	4.30	34	1.26	
Totals	19.81	39		

* Significance level of .05.

As the experiment was structured to hold comprehension quotients constant, no analysis of comprehension scores is presented.

DISCUSSION

The results of this study confirm the Whitehill and Jipson findings of the overall superiority of the operant or instrumental versus the traditional strategy of reading instruction. That is the instrumental method produced greater proportional WPM gains across E, I, and M groups than did the traditional methodology. These results do not entirely replicate the Whitehill and Jipson finding that E's do significantly better in the instrumental program than do I's or M's. Although E's did show more gain in the instrumental group than any other group in any other treatment, the overall difference is not great enough to reach a significant alpha level of the program effect.

The next phase of this research program will involve assessment of the instrumental versus the traditional program over a large number of Ss using a correlational design rather than experimental treatment. If reasonable cross validation is found, there would seem to be good reason to adopt the instrumental approach to speed reading instruction in general.

REFERENCES

1. Brown, James I., Efficient Reading, Heath and Company, Boston, Massachusetts, 1962.

2. Miller, Lyle L., Increasing Reading Efficiency, Holt, New York, New York, 1964.

3. Miller, Lyle L., Maintaining Reading Efficiency, Developmental Reading Distributors, Laramie, Wyoming, 1966.

4. Whitehill, Richard P.; Jipson, Janice A., "Differential Reading Program Performance of Extraverts and Introverts," The Journal of Experimental Education, 38:93-96, Spring 1970.

TWO GENERALIZATIONS OF THE ITEM

DISCRIMINATION INDEX

TO MULTI-SCORE ITEMS

DOUGLAS R. WHITNEY and DARRELL L. SABERS
University of Iowa   University of Arizona

ABSTRACT

Two general item analysis indices which apply to multi-score items are developed as generalizations of a popular index applicable to dichotomous items. The indices of discrimination are of two types: one based on differential difficulty and the other on net number of positive discriminations. The usefulness and limitations of each are discussed.

IN THE PAST 40 years many articles and books have been published in which item analysis techniques have been described and investigated. Most of these methods for evaluating test items and identifying items needing improvement have been developed for dichotomous test items. This emphasis on items which can be scored as correct or incorrect has perhaps grown out of the computational ease offered by such items and the early adoption of objective items by large testing programs. Tests composed of multi-score items (e.g., essay tests) are still extensively used by classroom teachers, but little attention has been given to making available corresponding techniques for item selection and improvement.

The purpose of this paper is twofold. First, two approaches to the concept of item discrimination will be reviewed. Second, methods for quantifying these generalized concepts, which are appropriate for all types of items, will be described. The development of these techniques is such that they are direct extensions of an index commonly used for dichotomous items.

DISCRIMINATION: TWO APPROACHES

Discrimination as Differential Difficulty

One common definition of the discriminating power of an item when it is used with a pair of examinee populations is the difference between item difficulties for the populations. Many such pairs of populations and many degrees of discrimination are

thus possible for an item. In practice, one usually obtains a single index of discrimination by estimating the population difference for "high" and "low" groups chosen on some criterion measure (e.g., total test score, cumulative grade point average). Although not specifically stated, Johnson (4) presumably used this approach in defining his U-L Index of item validity. His index depends on the difference between the proportion of correct responses (or equivalently, mean item score) for two examinee groups. In this context, positive discrimination refers to the degree to which the "high" examinee group attains a higher average score on the item than the " low " group. Presumably, the criterion on which discrimination is desired is not available for all individuals of interest to the examiner. For example, total test score is often used as substitute for a theoretical construct measure. High positive item discrimination suggests that the item will serve as a partial substitute for this criterion and rank Ss approximately as they would be ranked on the criterion.

In order to quantify this concept, it is necessary to first define item " difficulty " in the multi-score case. A general definition of difficulty is how "hard" the item is. That is, how does the performance on an item by a group of examinees compare with the highest possible level of performance. The performance level for a group may be expressed as the difference between their average item score and the minimum possible score on the item. The highest possible level of performance, expressed in a similar manner, would be the difference between maximum

and minimum possible item scores. The generally accepted convention of expressing difficulty as a percent will be used for this index. Thus, a general index of item difficulty is

$$P = [(\overline{X} - X_{min}) / (X_{max} - X_{min})] \times 100 \quad (1)$$

where P is the item difficulty in percentage units, $\overline{X}$ is the mean item score of n examinees, X_{max} is the highest obtainable item score, and X_{min} is the minimum obtainable item score. P represents the percent of maximum performance achieved by the group. The range of P is from 0 percent (when all Ss earned the minimum item score) to 100 percent (when all Ss earned the maximum item score). For items with a minimum score of zero, the index simplifies to $[\overline{X} / X_{max}] \times 100$, and becomes the percent of examinees answering the item correctly for a dichotomous item. When the traditional correction for guessing formula is applied to a multiple choice item, $X_{min} = -1/(r-1)$, where r is the number of item responses.

Suppose a multi-score item has been completed by twelve students, and that the scoring provided for integer scores of 0 to 3 inclusive. The following data might have resulted:

Item Score (X)	Number Earning Score (f)	Total Points by f Students fX
3	3	9
2	2	4
1	2	2
0	5	0
	N = 12	ΣfX = 15

The average item score ($\overline{X}$) is then $\Sigma fX/N = 1.25$. For this item X_{max} is 3 and X_{min} is 0, so that $P = [(1.25-0)/(3-0)] \times 100$, or about 42 percent. That is, this group earned 42 percent of the available points on this item.

The interpretation of item discrimination as differential difficulty leads to the definition of item discrimination (D_1) as the difference between item difficulties (as defined above) for two examinee groups. Since many previous discrimination indices have been expressed as decimal fractions ranging from -1 to +1, that convention will be used here. This index representing differential difficulty has the form:

$$D_1 = P_U - P_L = (\overline{X}_U - \overline{X}_L) / (X_{max} - X_{min}) \quad (2)$$

where P_U and P_L are the difficulties (as in formula (1)) for the upper and lower groups expressed as

proportions, $\overline{X}_U$ and $\overline{X}_L$ are the mean item scores for the upper and lower groups, and X_{max} and X_{min} are as previously defined.

Illustration of the Computation of D_1

Suppose that a multi-score item with possible scores 0 to 3 inclusive had been completed by twelve examinees in each of two criterion groups. The following data might have resulted:

Item Score (X)	Number in Upper Group Earning Score f_U	Total Points Earned $f_U X$	Number in Lower Group Earning Score f_L	Total Points Earned $f_L X$
3	5	15	1	3
2	4	8	1	2
1	2	2	4	4
0	1	0	6	0
	$n_U = 12$	$f_U \overline{X} = 25$	$n_L = 12$	$f_L \overline{X} = 9$

For this example, $\overline{X}_U = 25/12 = 2.083$, $\overline{X}_L = 9/12 = 0.75$, $X_{max} = 3$ and $X_{min} = 0$ so that $D_1 = (2.083 - 0.750)/3$ or about .44. That is, the upper group earned about 44 percent more available points than did the lower group.

Discrimination as Item-Criterion Association

An alternative definition for discrimination is the degree to which item performance is consistent with criterion performance. That is, the extent to which individuals who differ on the criterion measure differ in a similar manner on their item response. In most testing situations, item performance is expected to be related positively to criterion measures (e.g., total test score in the usual case). The degree of consistency of these performances could be demonstrated via some suitable correlation-type index. An alternate operational definition of this property is the net proportion of all possible subject pairs which show a positive relationship between item score and criterion measure. This approach was introduced by Findley (2) as an alternative rationale for the U-L index. Although the two approaches yield equivalent values for dichotomous items, generalization of Findley's approach leads to an index different from D_1 for multi score items.

For any number (N) of examinees, the number of possible unique pairs of Ss is $N(N-1)/2$. If these Ss are grouped into c criterion groups (of sizes n_1, n_2, ..., n_c; $\sum_{j=1}^{c} n_j = N$), those individuals within the same criterion group can no longer be differentiated. That is, since their criterion measures are the same, there can be no basis for discriminating among Ss within the same criterion group. Hence, these pairings cannot evidence either positive or negative

association with item response. Subtracting these pairs, or $\sum_{j=1}^{c} (n_j (n_j -1)/2)$ from the total number of possible pairs gives the maximum number of pairs which could show positive item-criterion association. It can be shown that $N(N-1)/2 - \sum_{j=1}^{c} (n_j (n_j - 1)/2)$ is algebraically equivalent to both $1/2 (N^2 - \sum_{j=1}^{c} n_j^2)$ and $\sum_{i=1}^{c-1} \sum_{j=i+1}^{c} n_j n_i$. The latter formula is usually preferable for computational purposes. This quantity, the maximum of positively discriminating pairs for fixed criterion group sizes, will be denoted by D_{max}.

The number of net positive discriminations may be determined by considering each of pairings involving one S from a criterion group and one S from another group. Each pair in which the S from a higher criterion (e.g., higher test score) group has a higher item score represents a positive (or correct) discrimination, and is counted + 1. Each pair in which the S from a lower criterion group has the higher item score is a negative (or incorrect) discrimination, and is counted -1. When Ss from different criterion groups have equal item scores, there is no discrimination and these pairs are not counted. The computation is simplified if the item data is put into matrix form as in the following example, and some intermediate sums are calculated. Here the index is

$$D_2 = (D_+ - D_-) / D_{max} \quad (3)$$

where D_+ represents number of positive associations, D_- represents the number of negative associations, and D_{max} represents the maximum number of positive associations.

Illustration of the Calculation of D_2

If item data are put into matrix form, with the ordered criterion groups as columns (column 1 is highest and column c is lowest) and ordered item sources as rows (row 1 is highest item score and row r is lowest—usually zero), then D_+ is the sum of the products of each cell frequency multiplied by the total number in all cells below and to the right of that cell. Similarly, D_- is the sum of the products of each cell frequency multiplied by the total number in all cells below and to the left of it. Computation of D_{max} is simplified if each criterion group contains the same number of examinees, n, for then $D_{max} = n^2 c(c-1)/2$.

For dichotomous items, D_+ is simply the product of the number of examinees in the upper group who got the item correct and the number of examinees in the lower group who got the item incorrect. Similarly, D_- is the product of the number of upper group incorrect responses and lower group correct responses. D_{max} is simply n^2, where n (as above) represents the number of examinees in each group. Only for items scored 0 or 1 and using 27 percent criterion groups is this index equivalent to Johnson's U-L index.

Again, the data below might have occurred for an item with scores from 0 to 3 inclusive. In this example, there are three ordered criterion groups, each containing ten Ss (although equal group sizes are used in all examples, neither index requires equal n's).

Item Score	Frequency Upper	Frequency Middle	Frequency Lower
3	5	3	0
2	3	3	2
1	2	3	4
0	0	1	4
Total	10	10	10

Here,

$$D_+ = 5(3+3+1+2+4+4) + 3(3+1+4+4) + 2(1+4)$$
$$+ 3(2+4+4) + 3(4+4) + 3(4) = 197$$
$$D_- = 0(3+3+1+3+2+0) + 2(3+1+2+0) + 4(1+0)$$
$$+ 3(3+2+0) + 3(2+0) + 3(0) = 37, \text{ and}$$
$$D_{max} = c(c-1) n^2/2 = 3(2)(100)/2 = 300 \text{ since the}$$

criterion groups each contain the same number of examinees. $D_2 = (197-37)/300$ or about .53 for this example. That is, of the three hundred pairs of Ss from different criterion groups, 53 percent more showed positive item-criterion association than showed negative association. Specifically 65.7 percent were positive and 12.3 percent were negative (the rest being neutral or non-discriminating).

DISCUSSION

These indices (D_1 and D_2), when applied to dichotomous items and two criterion groups, are equivalent to well-known indices. In fact, the discrimination indices D_1 and D_2 are identical for this type of item. However, this equivalence does not hold in general for other item types, and most of the discussion in this section concerns similarities and differences between the two indices.

Discrimination as Differential Difficulty (D_1)

The major advantage of this approach appears to be the conceptual and computational simplicity of index D_1. Since the concept of difficulty as described here can be grasped quickly, the index D_1 may be preferred for classroom teachers who are relatively new at using item analysis procedures. In addition, the computational ease recommends its use for teachers who do not have access to a desk calculator or other mechanical aids.

There is a limitation on the use of D_1 for item analysis work. This index is defined only for the case of two criterion groups. The use of two groups, particularly if they are extreme groups, may have certain advantages for the generalization of results (see below). However, it may be desirable to have a discrimination index to use with more than two criterion

groups. It would be possible to use some weighting of criterion group mean scores and thus obtain an index of differential difficulty similar to D_1, but much of the conceptual simplicity would have been introduced. A second limitation, the fact that D_1 is score-related, is discussed below.

Discrimination as Item-Criterion Association (D_2)

The major advantages of index D_2 as a measure of discrimination are almost exactly the weak points of D_1 and vice versa. First, D_2 is independent of the score values assigned to item responses. Specifically, the scores attached to the levels of item achievement have no effect on the value of D_2. This characteristic is appealing in that D_2 is dependent on the ability of the item to distinguish between achievement levels and is not affected by the scores attached to these levels.

A second advantage of D_2 is that it may be easily applied to situations which are not generally interpreted as test " items." An example of this use might be for aptitude test scores (where performance on the test is considered as the " item") used to distinguish between levels of success on a job. Another example might be in analyzing the degree of agreement between two raters or judges.

A third advantage of D_2 is that it is closely related to Kendall's (6) rank correlation tau. D_2 is, in fact, the product of tau and a function of the obtained response and criterion frequencies. Although not identical to Kendall's coefficient, D_2 is related to Cureton's (1) and Glass' (3) rank-biserial correlations. Further, an approximate (asymptotic normal) significance test and a randomization test based on Kendall's S (here $S = D_+ - D_-$) are available.

One limitation to the use of D_2 is that the use of more criterion groups than there are item response categories necessarily limits D_2 to some value less than unity. In the following example, the prescribed procedure for obtaining D_{max} indicates 540 possible positive pairs. Since a positively (or negatively) discriminating pair requires differing criterion and response categories, however, the maximum number of pairs with differing responses would be (18) (18) = 324. The limit on D_2 for this item would be $|D_2| \leq 324/540 = .60$.

Item Score	Frequency Upper Sixth				Frequency Lower Sixth	
1	6	6	0	0	0	18
0	0	0	6	6	6	18
Total	6	6	6	6	6	

In general, for an even number of criterion groups of equal size, $|D_2| \leq c (r-1)/ r (c-1)$ for $r \leq c$. The limiting value for an odd number of criterion groups is somewhat less than this value because of the necessity of splitting the middle criterion group frequency in order to achieve the maximum number of discriminating pairs allowed by r response groups. When r < c, the use of c criterion groups is, in effect, asking the item to make finer discriminations than its response categories will allow. Although D_2 may still be used in these cases, the values will generally be low when compared to cases for which c ≤ r.

Generally, the greater the number of criterion groups employed, the finer the discriminations examined in D_2. That is, the inclusion of the intermediate group in the example made a lower value of D_2 more likely (as compared to $D_2 = .78$ if only the upper and lower thirds had been used) because it introduced a group with a larger probability of misclassification. For this reason, obtained values of D_2 are dependent somewhat on the number of criterion groups employed and should only be compared with D_2 indices for items using similarly defined criterion groups. It has been shown (5) that using extreme groups minimizes the possibility of criterion misclassification relative to group size.

SUMMARY

D_1 and D_2 do not necessarily give identical values for the same data, nor do they yield the same information. There is no reason to expect identical results except that both reduce to the conventional U-L index in the case of dichotomous items and two criterion groups.

The D_1 index is easily computed and understood, but its value is dependent on the item scores employed. D_2 is perhaps applicable to a wider variety of testing situations, but is computationally complex. Both indices are related to the slope of the item characteristic curve for the item and criterion employed. D_1 is a direct estimate of the linear slope of this curve, between the two criterion subpopulation means. D_2, as a measure of item-criterion association, is also related to the item-characteristic curve. Its exact relationship is not yet fully explored.

A study to determine which of these (and other alternative) indices is most useful for certain purposes is underway. Specifically, their usefulness for selecting maximally reliable and/or valid subsets of items from an item pool will be investigated for items which are 6-point ratings. Also, comparisons between D values for items scored with and without a correction for guessing will be made, and computer simulation of essay item scores will be used to estimate the sampling distributions of the indices.

REFERENCES

1. Cureton, Edward, E., " Rank-biserial Correlation," Psychometrika, 21:287-290, September 1956.

2. Findley, Warren, G., "Rationale for the Evaluation of Item Discrimination Statistics," Educational and Psychological Measurement, 16:175-180 Summer, 1956.

3. Glass, Gene, V., " A Ranking Variable Analogue of Biserial Correlation: Implications for Shortcut Item Analysis," Journal of Educational Measurement, 2:91-95, June 1965.

4. Johnson, A. Pemberton, " Notes on a Suggested Index of Item Validity: The U-L Index," Journal of Educational Psychology, 42: 499-504, December, 1951.

5. Kelley, Truman, L., " The Selection of Upper and Lower Groups for the Validation of Test Items," Journal of Educational Psychology, 30: 17-24, January 1939.

6. Kendall, Maurice, G., Rank Correlation Methods Griffin Co., London, England, 1948, 160 pp.

BOOK REVIEWS

be said by Cronbach that he has not already covered carefully in previous editions? Perhaps surprisingly, a great deal. The largest of the standard measurement texts has become a good deal larger. Despite the somewhat smaller print that currently is in vogue, the new edition is one hundred pages longer than the 1960 edition (and that was a great deal longer than the first edition). The added length allows for expanded discussion of points briefly mentioned previously as well as giving attention to such "new" areas as "the computer as tester" and "testing the disadvantaged child." Again heavily documented, Cronbach has added over five hundred references that bear a date of 1960 or later and has nearly doubled the total number of references. The number of tables and figures has increased only slightly. More importantly, most have been replaced or updated. It is a dual tribute to the improved art work of test makers and the selection of Cronbach that, after looking at the slick, modern figures in the 1970 revision, the wooden stoves and expressionless people of the 1960 edition look like something lifted from a 1902 Sears catalog.

The title is accurate. The text's basic purpose remains unchanged. It presents the essentials—the method of inquiry, the critical standards, and the key concepts of the field. In his preface Cronbach states, "This book is intended to establish a base from which his (the reader's) knowledge can grow." He stresses that new thinking will be based upon the work already done, and although the reader should continue to learn from the unfolding research literature, established tests will continue to be used and updated to serve traditional functions.

Prefaces are so often filled with superficial chatter and strained acknowledgements that readers have learned to flip past them. In Cronbach's third edition this would be a gross error. He has used the occasion to review the societal conditions that influenced the course of education during the tenure of his three editions and the resultant influence upon tests, testing, and test theory.

Although he acknowledges that technical reporting on tests today far surpasses that of 10 or 20 years ago, his overall impression is that today's tests are obsolescent. While he acknowledges that the tests that were born before 1949, more often than not, are the best that we have today, he seems impatient for a breakthrough that will carry the field of measurement to a new level of sophistication. He laments that perhaps "the things actuarially scored tests cannot do are more important than the things they can do," and is "more convinced than ever that the solution for most of the ills of testing is to develop sound knowledge of aptitude-treatment interactions."

As one reads the material, the subheads of 1970 are both more frequent and more on-target. In fact, he introduces some eye-catchers such as, "The Bandwidth-fidelity Dilemma," "How Permissible is Deception?" "Obsolescence of Norms," and "The Signal-noise Ratio." Among the new sections that have been added are: "Evolution of the Testing Enterprise," "Expectation of Failure," and "Testing in Developing Nations." He has kept pace with the times by replacing a 1960 section entitled "Development of a Stenographic Aptitude Test" with "Development of an Aptitude Test for Computer Programmers."

Sprinkled throughout the volume are subtle indications that, if the author does not view the world of measurement as a more complex venture than he did a decade ago, he has elected to speak in a more tentative voice and offer fuller explanation of his position. The authoritarian "How to Choose Tests" has been tempered to a more mellow "Other Characteristics Desired in Testing" both through the change in title and through a more careful consideration of sources of error. "Personality Measurement" no longer has the meter-stick precision of sound. Although the material is largely the same, the complexity of the subject is conveyed through Cronbach's more cautious re-titling of the chapter as "Studying Personality." Whereas he could formerly present "The Self-description as a Report of Typical Behavior," he now more questioningly has titled the section, "The Self-description: Report of Typical Behavior?"

The amount of attention given to the IQ has been sharply reduced. Cronbach has concentrated, instead, upon research and theory concerning ability tests and devotes sections to the work of Piaget, cultural influences, and tests for infants and preschoolers. Also, he has devoted given detailed attention to Guilford's work on convergent and divergent thinking, the social intelligence hypothesis, and a simulation model for bead-chain performance. Surprisingly, discussion of proficiency tests has been eliminated, not because of lack of current work in the field, but rather because the field has exploded beyond the bounds of treatment in a general text.

Cronbach's usual high standard of scholarship is evident throughout. Despite his lamentations about the static condition of the field, a comparison of the second and third editions makes it evident that the 1960 edition has been obsolesced. Reader, take note.

Professor Norman R. Stewart, Reviewer
Michigan State University

A MULTIPLE REGRESSION APPROACH

TO MULTIPLE COMPARISONS FOR COMPARING

SEVERAL TREATMENTS WITH A CONTROL

JOHN D. WILLIAMS
The University of North Dakota

ABSTRACT

A multiple regression approach is used for the multiple comparison situation in which several treatment groups are compared to a control group. Using a regression approach, it is shown that the multiple comparison procedure described by Dunnett can be found as a by-product of the general regression program; the test described by Dunnett yields identically the same results as the test of significance for the partial regression weights.

RECENT efforts (notably Bottenberg and Ward (1) and Jennings (5)) have been made to present multiple linear regression as a problem-solving technique. Ward (8) has compared four different approaches to problem solving: analysis of variance, multiple regression, analysis of covariance, and a technique called VARICO - a " sort of reverse covariance analysis. " Ward showed that, while the methods differed conceptually, the four approaches have many basic ideas in common. The difficulty of recognition of this situation, i. e., the relationship that exists between the usual analysis of variance approach and multiple regression as a general data-analytic system has been discussed by Cohen (2). Jennings (5) discussed at length a 2x3 fixed effects analysis of variance from a regression viewpoint. Both Jennings and Ward have extensively used a binary coding to effect a solution.

One criticism that has been made of multiple regression approach, as compared to the analysis of variance approach, is that while the analysis of variance can be duplicated by a regression analysis, no real advantage is gained. Without discussing this criticism in detail (Cohen (2) has already done so), this article shows an additional conceptual usefulness of a multiple regression approach by focusing on a particular application of a regression approach to the problem of multiple comparisons. The regression approach is conceptually simple and yields a multi-

ple comparison procedure for comparing several treatments with a control, often referred to as Dunnett's (3, 4) test. It should be made clear from the onset that the present effort does not purport to extend Dunnett's test, but rather, using multiple linear regression, to obtain exactly the same results with considerably less effort. It also lends some meaning to the testing of significance for the partial regression weights.

To show the relationship between a regression approach and the usual analysis of variance approach, an example using sample data is first subjected to the computations of analysis of variance, and then Dunnett's test is run. The problem is then reformulated from a regression viewpoint, and finally comparisons of the two methods can be made.

AN EXAMPLE

The following data are presented for analysis:

Control Group	Group I	Group II	Group III
9	8	13	15
8	7	10	12
6	8	12	10
3	6	11	17
4	6	14	11
$\overline{X}_0 = 6.0$	$\overline{X}_1 = 7.0$	$\overline{X}_2 = 12.0$	$\overline{X}_3 = 13.0$

TABLE 1

SUMMARY TABLE FOR COMPARING SEVERAL TREATMENTS WITH A CONTROL

Source of Variation	df	Sum of Squares	Mean Squares	F
Among groups	3	185.00	61.677	13.333
Within groups	16	74.00	4.625	
Total	19	259.00		

The first group has been labeled X_0 and serves as the control group to which all other groups are compared. As a first step, a summary table (See Table 1) is presented using the usual analysis of variance approach.

Dunnett's test is a test to compare p treatments with a control. The means are given by $\overline{X}_0$, $\overline{X}_1$, ... $\overline{X}_p$, where the $\overline{X}$'s are means of p + 1 sets of observations which are assumed to be independently and normally distributed in the population. $\overline{X}_0$ refers to the control group, and each $\overline{X}_i$ (i = 1, ..., p) refers to the ith treatment group mean. Dunnett's original article (3) assumes an equal number of observations in each set, and tables are presented to test for significant differences between each $\overline{X}_i$ and $\overline{X}_0$. Both one- and two-sided tables are provided; however, Dunnett's second article (4) should be consulted for two-sided tests because of computational approximations for the case of more than two comparisons in the original article.

Dunnett's test is given by

$$t = \frac{\overline{X}_i - \overline{X}_0 - (m_i - m_0)}{s\sqrt{\frac{1}{N_i} + \frac{1}{N_0}}} \quad (1)$$

Using MS_w as the estimate of s^2, and assuming all groups are of equal size, with the hypothesis $m_1 - m_0 = 0$, then (1) can be reduced to

$$t = \frac{\overline{X}_i - \overline{X}_0}{\sqrt{\frac{2(MS_w)}{n}}} \quad (2)$$

Dunnett's original article develops confidence limits for each comparison; a critical difference approach can also be used so that the comparisons can be made quickly:

$$c.d. = t_i\sqrt{\frac{2(MS_w)}{n}} \quad (3)$$

where t is taken from tables prepared by Dunnett (3, 4). While the critical difference methodology is quite convenient, for this presentation a t value is found for each comparison with the control for the data in Table 1, so that the present regression approach and the usual analysis of variance approach can be compared.

The three comparisons to the control group can be effected by equation (2).

$$t_1 = \frac{7.0 - 6.0}{\sqrt{\frac{2(4.625)}{5}}} = .735$$

Similarly, $t_2 = 4.411$ and $t_3 = 5.147$. Using Dunnett's tables, t_2 and t_3 are both significant at the .01 level, while t_1 is not significant.

A REGRESSION APPROACH

On the other hand, the problem can be viewed from a regression viewpoint. It is helpful to define four binary predictors:

$X_0 = 1$ if the score is from a member of the control group; and 0 otherwise

$X_1 = 1$ if the score is from a member of group 1; and 0 otherwise

$X_2 = 1$ if the score is from a member of group 2; and 0 otherwise

$X_3 = 1$ if the score is from a member of group 3; and 0 otherwise

A linear model can be written for this situation:

$$Y = b_0 + b_1X_1 + b_2X_2 + b_3X_3 + e \quad (4)$$

where

b_0 = the Y-intercept
b_1 = the regression coefficient for group 1
b_2 = the regression coefficient for group 2
b_3 = the regression coefficient for group 3
e = the error involved in prediction

TABLE 2

REGRESSION FORMULATION FOR COMPARING SEVERAL TREATMENTS WITH A CONTROL

Y	X_0	X_1	X_2	X_3
9	1	0	0	0
8	1	0	0	0
6	1	0	0	0
3	1	0	0	0
4	1	0	0	0
8	0	1	0	0
7	0	1	0	0
8	0	1	0	0
6	0	1	0	0
6	0	1	0	0
13	0	0	1	0
10	0	0	1	0
12	0	0	1	0
11	0	0	1	0
14	0	0	1	0
15	0	0	0	1
12	0	0	0	1
10	0	0	0	1
17	0	0	0	1
11	0	0	0	1

TABLE 3

OUTPUT OF MULTIPLE REGRESSION PROGRAM

Variable No.	Mean	Standard Deviation	Correlation X vs Y	Regression Coefficient	Std. Error Of Reg. Coef.	Computed t Value	Beta
3	0.25000	0.44426	-0.40109	1.00001	1.36014	0.73522	0.12033
4	0.25000	0.44426	0.40109	6.00001	1.36014	4.41130	0.72197
5	0.25000	0.44426	0.56153	7.00001	1.36014	5.14652	0.84230
Dependent 1	9.50000	3.69210					

Intercept   5.99999

Multiple Correlation   0.84515

Std. Error of Estimate   2.15058

Analysis of Variance for the Regression

Source of Variation	Degrees Of Freedom	Sum of Squares	Mean Squares	F Value
Attributable to Regression	3	185.00027	61.66675	13.33340
Deviation from Regression	16	73.99973	4.62498	
Total	19	259.00000		

It can be noticed that the control group has seemingly been left out. However, if this equation is solved for an expected value for a member of the control group,

$$E(Y) = b_0 + b_1(0) + b_2(0) + b_3(0)$$

$E(Y) = b_0$. The expectancy for a member of the control group will by definition be $\overline{X}_0$. Thus, a least squares solution for b_0 is $\overline{X}_0$, the mean of the control group.

For a member in group 1, the expected value would be

$$E(Y) = b_0 + b_1(1) + b_2(0) + b_3(0)$$
$$E(Y) = b_0 + b_1$$
$$E(Y) = \overline{X}_0 + b_1. \qquad (5)$$

A least squares solution for the expectancy of a given member of group 1 is the mean of group 1;

Thus

$$\overline{X}_1 = \overline{X}_0 + b_1 \qquad \text{from equation (5), or}$$

$$\overline{X}_1 - \overline{X}_0 = b_1. \qquad (6)$$

Likewise

$$b_2 = \overline{X}_2 - \overline{X}_0 \text{ and } b_3 = \overline{X}_3 - \overline{X}_0.$$

Equation (4) can be rewritten

$$Y = \overline{X}_0 + (\overline{X}_1 - \overline{X}_0)X_1 + (\overline{X}_2 - \overline{X}_0)X_2 + \qquad (7)$$
$$(\overline{X}_3 - \overline{X}_0) X_3 + e.$$

Equation (7) lists precisely the comparisons of interest for comparing several treatments with a control. Since equation (4) (and, therefore, equation (7)) is the full model for the expression of a one-way analysis of variance, this approach also yields results identical to the analysis of variance situation. Thus, using equation (4), it can be seen that these two useful results can be obtained simultaneously: the usual analysis of variance as one part of the output, and Dunnett's test as the other part.

The information necessary for a regression solution, with equation (4) as the linear model, can be conveniently placed in tabular form (see Table 2).

For the data in Table 2, the general purpose multiple regression program was used. Table 3 contains the printout from that analysis. The variable number in Table 3 refers to the order in Table 2; the criterion variable is variable number 1; variable 2 refers to the control group, variable 3 to group 1, variable 4 to group 2, and variable 5 to group 3. Because variable 2 refers to the Control Group, no information appears in the printout using this variable number. The table of residuals has not been included herein.

Table 3 contains the previously mentioned items. It can be recalled that $\overline{X}_0 = 6.0$, $\overline{X}_1 = 7.0$, $\overline{X}_2 = 12.0$, $\overline{X}_3 = 13.0$. The intercept is 6.0 (within rounding error) and is $\overline{X}_0$. Also $b_1 = 1 = \overline{X}_1 - \overline{X}_0$, and is

in keeping with equation (7). Similar statements could be made concerning b_2 and b_3. Of more interest for this particular presentation is that the computed t values are identical (to three decimal places) to the t values found by using equation (2), which is Dunnett's test for equal-sized groups.

Dunnett's (3, 4) tables are necessary for preserving the probability level. If the experimental groups are not of equal size, then the test described by Dunnett, and therefore the present formulation, results in an approximate test.

DISCUSSION

A major reason for using multiple comparison procedures has been to make individual comparisons of means and to simultaneously preserve the probability level. One of several multiple comparison procedures that has emerged in the last 20 years has been Dunnett's test for a control. The major impetus of the present article has been to greatly simplify the process of obtaining this multiple comparison method. This simplification has been accomplished through the use of a multiple linear regression approach; this particular approach is quite simple and direct and with the use of Dunnett's tables, will preserve the probability level.

On the other hand, other ways of using multiple linear regression (7) will yield similar results. As an example of McNeil and Kelley's approach, the comparison of group 1 and the control group can be considered. The full model could be expressed as

$$Y = b_0 U + b_c X_c + b_1 X_1 + b_2 X_2 + b_3 X_3 + e_1 \quad (8)$$

The model given in equation (4) and equation (8) have the same notation, with these exceptions:

U = a unit vector (i.e., a predictor vector containing 1's).
b_c = the regression coefficient for the control group
X_c = the control group
e_1 = the error involved in prediction for equation (8)

To test the hypothesis that the control group and group 1 are equal, the following restriction can be made:

$b_c = b_1$ (This is the same hypothesis as $\overline{X}_c = \overline{X}_1$).

Then equation (8) can be rewritten:

$$Y = b_0 U + a_0 (X_c + X_1) + b_2 X_2 + b_3 X_3 + e_2 \quad (9)$$

where

a_0 = the regression coefficient for the combined groups of X_c and X_1
e_2 = the error involved in prediction for equation (9)

Comparing equation (9) to equation (8) in the methodologies of Kelley and others (6), identically the same result ($F = .5405 = (.735)^2$) is achieved as is in Table 1. Similar procedures would yield the other comparisons.

SUMMARY

The present paper has presented a specific appli-

cation of multiple regression as a problem-solving technique to the problem of multiple comparisons. Results of an analysis of variance and the subsequent multiple comparisons of several treatments with a control are given. When using the regression approach presented herein, those same comparisons can be read directly from a regression printout, which illustrates that Dunnett's test can be conceptualized as a test of significance for a partial regression weight. This is accomplished by setting the regression coefficient for the control group equal to zero in a linear model. This is done by not including the vector for the control group in the prediction equation (linear model). Effectively, if the researcher wishes to make comparisons of several treatments with a control, he needs only to binary code the group membership and use the resulting binary coded vectors (not including the control group binary coded vector) as predictors (as demonstrated herein). The computed t test for each partial regression weight is identical to the test Dunnett suggested for comparing several treatment groups to a control group. Thus, no additional computations need be made, either by a calculator or by additional computer runs for this situation.

REFERENCES

1. Bottenberg, R. A.; Ward, J. H., Jr, Applied Multiple Linear Regression, Personnel Research Laboratory, Aerospace Medical Division, Lackland Air Force Base, Texas, (PRL-T D R - 63 6), 1963.

2. Cohen, J., "Multiple Regression as a General Data-Analytic System," Psychological Bulletin, 70: 426-443, 1968.

3. Dunnett, C. W.; "A Multiple Comparison Procedure for Comparing Several Treatments with a Control," Journal of the American Statistical Association, 50: 1096-121, 1955.

4. Dunnett, C. W., "New Tables for Multiple Comparisons with a Control," Biometrics, 20: 482-491, 1964.

5. Jennings, E., "Fixed Effects Analysis of Variance by Regression Analysis," Multivariate Behavioral Research, 2: 95-108, 1967.

6. Kelley, F. J.; Beggs, D. L.; McNeil, K. A.; Eichelberger, T.; Lyon, J., Multiple Regression Approach, Southern Illinois University Press, Carbondale, Illinois, 1969.

7. McNeil, K. A.; Kelley, F. J. personal communication, 1970.

8. Ward, J. H., Jr., "Synthesizing Regression Models—An Aid To Learning Effective Problem Analysis," The American Statistician, 23: 14-20, April 1969.

DIRECTIONS FOR J.E.E. CONTRIBUTORS

The Journal of Experimental Education publishes specialized or technical education studies, treatises about the mathematics or methodology of behavioral research, and monographs of major current research interest.

ABSTRACT REQUIRED

An abstract of not more than 120 words must accompany each manuscript. It should precede the text, be designated *ABSTRACT,* and conform to the following standards:

1. In a research paper, include statements of (a) the problem, (b) the method, (c) the data, and (d) the conclusions.

2. In a review or discussion article, state the topics covered and the central thesis.

3. Standard abbreviations may be used freely if the meaning is clear. Use complete sentences and do not repeat information which is in the title.

TEXT

Usually research articles should have a clearly organized order of presentation. The following elements and their order is a guide and is not intended to be prescriptive.

The Problem. The nature, scope, and significance of the problem should be presented.

Related Research. Presentation and discussion of related research should be selective and should include references essential to clarify the problem under examination.

Methodology. This section should consist of hypotheses, description of the sample and sampling procedures, discussion of the variables, description of the data-gathering instruments (if well-known, their description may be omitted), and general description of the statistical procedures.

Presentation and Analysis of Data. Analysis of the data and conclusions about the hypotheses should be more than mere presentation. Rival hypotheses and significance for related studies and educational theory should be considered. Summary data (means, standard deviations, frequencies, correlation coefficients, etc.) should be presented in tabular or graphic form. Discussion of these data should be interpretive.

Summarizing Statements. A summary of conclusions and implications for education may supplement the abstract.

STYLE

Certain suggestions are offered here to achieve uniformity in publication style and to conserve the time and energy of the author and editorial staff in the preparation of copy. *A Manual on Writing Research,* 1962, and *A Manual of Form for Theses and Term Reports,* 1962, by Kathleen Dugdale, the Indiana University Bookstore, Bloomington, may be used as style manuals in preparation of manuscripts.

Two Copies Required. Copies should be double spaced with wide margins. Typed copies are preferred, but dittoed or mimeographed copies will be accepted if they are legible.

Subheads. Articles are frequently improved by the judicious use of subheads. However, avoid use of the title, INTRODUCTION, for a lead section.

Title. Try to use a short title, preferably no more than ten words. Avoid superfluous phrases, such as "A Comparison of . . .," "A Study of . . .," and "The Effectiveness of . . . ."

Tables. Prepare tables precisely as they are to appear in *The Journal,* caption each with a brief and to-the-point title, and number consecutively with arabic numerals: *Table 3.* INTERCORRELATION MATRIX, VARIABLES OPTIMALLY ORDERED.

Figures. Draw any graphs and charts in black ink on good quality paper, title each, and number consecutively, thus: *Figure 4.* SCHOOL ENROLLMENT. Send the original copy. We cannot accept mimeographed figures or charts. Data should not be framed, and ordinates and abscissas should be shortened to occupy as little space as possible.

Tables and Figures. Tables and figures must be original copies acceptable for reproduction. A charge will be assessed for any redrawing or re-typing of tables or figures.

Technical Symbols. All symbols, equations, and formulas must be clearly represented. Differentiate between numbers and letters (zero and the letter o; one and the letter l, etc.) Identify hand-drawn Greek letters in the margin. Align subscripts carefully.

Footnotes. Avoid explanatory footnotes by incorporating their content in the text. However, for essential footnotes, identify them with consecutive superscripts, as book,[2] study,[3] etc., and list the footnotes in a section, entitled FOOTNOTES, at the end of the text, but preceding the REFERENCES.

References. References should be listed alphabetically according to the author's last name at the end of the manuscript. Each entry should be numbered. Citation of a reference in the text should be made with this number, as (2); or if specific pages are to be indicated, as (2:245–7).

Illustrations of article and book references:

1. Bittner, Reign H. and Wilder, Carlton E., "Expectancy Tables: A Method of Interpreting Correlation Coefficients," *The Journal of Experimental Education,* 14:245–52, March 1946.
2. Thomson, Godfrey, H., *The Factorial Analysis of Human Ability,* Houghton Mifflin Co., Boston, 1950, 383 pp.

COSTS

The publisher charges a contributor's fee of $6 per printed page of approximately 1,200 words, billed upon publication. Authors are charged for changes in tables, figures, or copy made when article is in camera-ready form. Each contributor will receive 10 complimentary copies of the issue in which his article appears. Reprints are charged at cost, and a price schedule will be sent to each contributor.

PROOFREADING

We will send you proofs for correction (with instructions for handling). Any major changes made in the proofs that were not incorporated in your original copy will be an added expense to you. (Errors that we make, naturally, will be at our expense.)

Keep an exact carbon copy of your manuscript so that if a question arises the editors can refer you to specific pages, paragraphs, or lines for clarification by letter.

SEND MANUSCRIPTS TO

John Schmid, Department of Research and Statistical Methodology, University of Northern Colorado, Greeley, Colorado 80631.

THE *Journal* OF
Experimental
Education

CONTENTS

Arvil S. Barr, Founder

EDITOR AND PUBLISHER ● 1932-1962

FACULTY ATTITUDES TOWARD UNIVERSITY

ROLE AND GOVERNANCE:

A FACTOR ANALYTIC APPROACH

L. ERWIN ATWOOD and KENNETH STARCK[1]
Southern Illinois University, Carbondale

ABSTRACT

Despite the changing role of the university faculty member (7), little inquiry has been made in t o faculty perceptions concerning university role and governance. Utilizing a 29-item instrument devised by the Educational Testing Service (ETS), this study sought (1) to determine what faculty members regard as some of the issues and (2) to identify any existing patterns of opinion. The sample consisted of 132 interviews with faculty members of a large Midwestern university. R-analysis isolated four dimensions of concern. Two centered on aspects of freedom and control within the university: the teaching versus research dichotomy, and university involvement in societal concerns. Q-analysis yielded three basic opinion types. These included concerns for academic freedom and control, social activism of the university, and research versus teaching. Multiple linear regression indicated that predominant demographic characteristics associated with opinion types w e r e political orientation, type of job, and length of time teaching.

EXAMINATION of attitudes of university faculty members toward various issues has been the legitimate concern of educational investigators for some time. Ranging from rather crude measuring devices yielding frequency data to elaborate instruments designed to utilize powerful statistical techniques, the studies generally deal with a single concern, such as classroom behavior (3) and liberal versus professional education (2). A few studies, including the Faculty Attitude Survey (9), have sought to chart the dimensions of "faculty morale" and "satisfactions and dissatisfactions" (10). Even fewer attempts have been made to delineate faculty perceptions of issues that have assumed paramount importance in higher education of the 1960's, namely, the definition of the role and regulation of the university.

Indications are that the rapid growth of the university has produced a change in the role of the faculty member. Generally, the change has diminished his autonomy and made it impractical for him to participate directly in policy making (7). Graybeal (4) reported results of a national survey in which college and university faculty members were asked about institutional practices involving promotions, academic freedom, and faculty authority; however, apparently no attempt was made to relate particular issues to one another or to identify attitude patterns among the respondents.

PURPOSE

The purpose of this study was to determine how faculty members perceive certain selected issues centering on the university's role in society and the regulation and control of the university as represented by a set of twenty-nine statements. Further, an attempt was made to explore the relationships between opinion patterns and such demographic chacteristics as age, political orientation, and type of job.

METHOD

Instrument

In early 1969 the ETS released results of a national study which dealt with college and university boards of trustees (8). Besides providing for data on the role of the trustee and personal background information, the questionnaire developed for t h e study included twenty-nine Likert-type statements designed to examine the trustees' perceptions of the role of higher educational institutions and their governance. Respondents were asked to check one of the five alternatives that best represented their feelings about each statement. The alternatives were: strongly agree, agree, don't know, disagree, a n d strongly disagree.

The first eighteen statements were to be regarded

by the respondents as applicable to their institutions; the remaining eleven statements were to be regarded in terms of higher education as a whole. These twenty-nine statements, with minor adaptations, together with twelve demographic items, comprised the instrument for the current study. The statements appear in Table 1.

Sample

A simple probability sample of 220 faculty members was drawn from the faculty directory of a Midwestern university which at the time of the study—Spring 1969—had an enrollment of about 20,500. Only the academic ranks of instructor, assistant and associate professor, and professor were included. The sample included a number of persons whose primary duties were not necessarily instructional inasmuch as nearly all administrative and many service personnel carry academic rank within one of nine academic units at the institution. Further, the faculty directory did not specify the individual's primary duties. The population numbered eight hundred. Of the 220 personal interviews assigned, 145 were completed; of these 132 questionnaires, or 60 percent of the original sample, were usable.

By academic unit there was little discrepancy between the number of respondents and non-respondents, e.g., 44.6 percent of the respondents were in the College of Liberal Arts and Sciences, which has 46.6 percent of the entire university faculty. Of the respondents, 60.7 percent held doctoral degrees compared with 53.3 percent for the population. By faculty rank, the respondents underrepresented professors and associate professors, 20.8 percent and 13.1 percent respectively, as compared with 33.0 and 27.5 percent for the population. Overrepresented among the respondents were assistant professors, 37.7 percent versus 31.8 percent for the population, and instructors, 28.5 percent versus 7.6 percent for the population.

As for the non-respondents, the percentage of women not responding, 22.1 percent, was higher than the percentage of women in the total population (15.4 percent). The number of associate professors and professors not responding, 20.0 and 28.4 percent respectively, was somewhat higher than for those from the total population responding, 13.1 percent for associate professors and 20.8 percent for professors.

Data Treatment

Factor analysis (1, 5) offered an appropriate means of determining (1) what issues respondents perceived within the universe of twenty-nine statements, and (2) what opinion patterns existed among the respondents. For the issues, this meant factor analysis of correlations between all pairs of statements (R-factoring); for opinion patterns, factor analysis of correlations of patterns of responses between all pairs of respondents (Q-factoring). An opinion pattern was defined as a hierarchy of the twenty-nine statements that represented the pattern a respondent felt was most representative of how he felt about the statements. Limitations of the computer program, which had a capacity of 109 variables and 130 observations, necessitated elimination from the analysis of data from several respondents. Thus, for the R-factoring, two of the 132 respondents in the sample were elimi-

nated at random; for Q-factoring, an additional twenty-one respondents were dropped in the same way with the restriction that none of the twenty-one have research as his primary responsibility. Only seven of the 132 respondents fell into the research category.

Multiple linear regression (6) afforded a means of evaluating relationships between opinion types and demographic characteristics. Factor loadings for each opinion type served as criteria, with demographic variables serving as predictors. In each case data were tested for curvilinearity and interaction effects before the linear models were analyzed; none were found in the models tested.

RESULTS

Issues

For the R-factor analysis a principal axis solution was used with rotation to orthogonal (Varimax) simple structure. The minimum eigenvalue for factoring was 1.0. Three distinct dimensions emerged with a fourth appearing to be related to one of the three. When a 3-factor solution was requested, Factors 1 and 4 collapsed into a single dimension without changing the interpretation of Factors 2 and 3. Correlations among all pairs of statements and the simple structure factor matrix are given in Tables 2 and 3. While Factor 1 represented a more generalized control dimension, Factor 4 was concerned primarily with faculty freedom and participation in the determination of university policy. For example, these two statements best represented Factor 4:

> There should be faculty representation on (name of university) governing board.

> (Name of university) faculty members should have the right to express their opinions about any issue they wish in various channels of university communication, including the classroom, student newspaper, etc., without fear of reprisal.

These two statements best represented Factor 1:

> The (name of university) administration should exercise control over the content of the student newspaper.

> Attendance at (name of university) is a privilege, not a right.

Factor 2 appeared to revolve around the seemingly endless conflict between those with teaching, as opposed to research, orientations. Although showing a slight concern for curriculum, Factor 2 seemed best represented by these two statements:

> The value of the PhD (EdD) is overemphasized in recruiting a faculty at (name of university).

> Teaching effectiveness, not publications, should be the primary criterion for promotion of faculty at (name of university).

Factor 3 was one of social concern. This dimension

TABLE 1

Z-SCORES FOR ALL OPINION TYPES FOR EACH OF THE TWENTY-NINE STATEMENTS*

No.	Statement	Opinion Type		
		I	II	III
1.	Attendance at (name of university) is a privilege, not a right.	-0.0	1.3	0.7
2.	In making admissions decisions, academic aptitude should be the most important criterion (i.e., given the greatest weight) at (name of university).	0.8	0.5	-0.8
3.	(Name of university) faculty members should have the right to express their opinions about any issue they wish in various channels of university communication, including the classroom, student newspaper, etc., without fear of reprisal.	1.7	-0.7	0.7
4.	The (name of university) administration should exercise control over the content of the student newspaper.	-1.7	0.6	-0.9
5.	All campus speakers should be subject to some official screening process.	-1.6	0.5	0.1
6.	There should be faculty representation on (name of university) governing board.	1.7	0.2	0.2
7.	Students who actively disrupt the functioning of (name of university) by demonstrating, sitting in, or otherwise refusing to obey the rules, should be expelled or suspended.	-0.3	1.4	1.2
8.	The grading system now in use at (name of university) needs to be modified.	0.4	-0.4	0.1
9.	An active research interest is a prerequisite for good undergraduate teaching. A man who does no research on a subject soon becomes less qualified to teach it.	0.4	-0.6	-2.0
10.	The value of the PhD or EdD is overemphasized in recruiting faculty at (name of university).	-0.8	-2.3	1.2
11.	(Name of university) should be actively engaged in solving contemporary social problems.	1.0	0.8	0.4
12.	Teaching effectiveness, not publications, should be the primary criterion for promotion of faculty at (name of university).	0.4	0.1	1.6
13.	(Name of university) should serve as a cultural center for the population in the surrounding area.	1.4	1.5	1.2
14.	(Name of university) curriculum should be deliberately designed to accommodate a wide diversity in student ability levels and educational-vocational aspirations.	0.6	0.9	0.6
15.	(Name of university) should be as concerned about the personal values of its students as it is with their intellectual development.	0.0	1.0	1.2
16.	Students involved in civil disobedience off the (name of university) campus should be subject to discipline by the college as well as the local authorities.	-1.6	-0.2	-1.3
17.	There should be more professional educators on (name of university) board of trustees.	0.6	-0.9	-0.3
18.	The more appropriate role of the (name of university) president is that of mediator rather than leader.	-0.5	-2.3	-2.1
19.	There should be opportunities for higher education available to anyone who seeks education beyond secondary school.	0.8	0.2	1.2
20.	The requirement that a professor sign a loyalty oath is reasonable.	-1.7	0.1	-0.9
21.	A definite institutional religious commitment does not necessarily preclude a genuine exposure of the students to alternative views nor prevent free inquiry and expression on the part of the faculty.	0.5	0.8	-0.3
22.	Increased federal support of higher education will mean increased federal control.	-0.3	-0.5	-0.7

TABLE 1   (Continued from previous page)

No.	Statement	Opinion Type		
		I	II	III
23.	The typical undergraduate curriculum has suffered from the specialization of faculty members.	-0.2	-1.4	0.4
24.	Colleges should admit socially disadvantaged students who do not meet normal entrance requirements.	0.8	0.2	0.1
25.	Traditionally Negro institutions serve a necessary function by offering the Negro student a curriculum which more nearly meets his needs and educational background.	-1.0	-1.2	-1.3
26.	A coeducational institution provides a better educational setting than a college for only men or women.	1.1	1.4	1.2
27.	Collective bargaining by faculty members has no place in a college or university.	-0.8	-0.2	-0.1
28.	Running a university is basically like running a business.	-1.4	-0.9	-0.9
29.	Fraternities and/or sororities or similar social clubs provide an important and positive influence for undergraduates.	-0.2	0.2	-0.5

* Statements from the College Trustee Study questionnaire. Copyright 1968 by Educational Testing Service. All rights reserved. Adapted by permission.

saw the grouping together of statements related to problems of current social conditions and, to a lesser extent, problems of curriculum revision and reorganization of the academic community. Factor 3 differed from Factor 2 on curriculum; the former was concerned with changing the curriculum to meet new demands, while the latter appeared more interested in pitting "teaching" against "research." The two statements best typifying Factor 3 were:

> (Name of university) should be actively engaged in solving contemporary social problems.

> Colleges should admit socially disadvantaged students who do not meet normal entrance requirements.

These four factors, although accounting for only 30.18 percent of the total variance, appeared to exhaust the dimensions the respondents preceived to exist within the set of twenty-nine statements (see Table 3). It should not be assumed that these four dimensions exhaust all possible meaningful factors that different sets of respondents might perceive or that might result from a different sampling of items.

Opinion Types

For the Q-analysis, the raw score matrix was normalized before correlations were computed. The factor analysis was a principal axis solution with rotation to orthogonal (Varimax) simple structure. The minimum eigenvalue criterion for factoring was 1.0. For each Q-type, that is, opinion type, standard scores (z-scores) were calculated for each of the twenty-nine statements according to procedures originally outlined by Stephenson (11). A z-score difference of ± 1.0 served as the criterion for a meaningful difference between opinion types on any statement.

Thus, z-scores for a given opinion type which were greater than +1.0 indicated strong agreement with the statement; z-scores less than -1.0 indicated strong disagreement. And where z-score differences across all types were less than ± 1.0, the statements were considered consensus items.

The Q-analysis isolated three basic patterns of opinions—accounting for 41.64 percent of the total variance—among the 109 respondents. [2] Again, other opinions may exist among these faculty members. Likewise, these three patterns do not necessarily encompass all members of the university faculty. On the other hand, the investigators are fairly confident that these patterns were the predominant patterns among most of the faculty at the time of the survey.

To begin with, it might be useful to summarize the points on which all respondents generally agreed. Twelve of the twenty-nine statements were consensus statements, and all twelve were concerned with socially oriented problems. This suggests there is less disagreement among faculty members over the third R-factor dimension—social concern—which is probably the most socially acceptable of the twenty-nine statements. The two consensus statements with which there was strong agreement were:

> (Name of university) should serve as a cultural center for the population in the surrounding region.

> A coeducational institution provides a better educational setting than a college for only men or women.

Neither statement appears to be the type that would arouse widespread argument. But this may not be true of the following statement with which there was strong disagreement.

TABLE 2

CORRELATIONS AMONG ALL PAIRS OF STATEMENTS *

	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29
1.	-00	-05	48	35	07	39	-12	-01	09	-15	25	07	14	32	33	-01	-25	-12	37	18	17	00	-20	14	31	36	34	20
2.		19	00	08	12	06	-04	22	-27	01	-14	04	01	-16	07	02	13	-11	-09	13	-04	-14	-16	02	-18	-01	-17	06
3.			-33	-27	33	-09	16	05	11	09	06	09	-05	-10	-22	18	21	-10	-36	-13	-10	-08	04	-02	04	-21	-22	-06
4.				57	-25	42	-16	01	04	-20	05	-08	15	37	44	-21	-29	-10	54	16	16	-07	-28	25	11	36	29	02
5.					-16	43	-20	-09	07	-21	15	-13	05	30	33	-12	-25	06	56	-00	24	07	-34	25	-03	29	27	07
6.						-14	10	15	05	22	02	02	08	-05	-03	31	20	12	-21	-06	-09	-07	05	-02	-06	-30	-22	-05
7.							03	-14	10	-20	31	06	12	37	37	-30	-36	-02	36	03	-07	10	-25	18	13	34	30	10
8.								-02	18	11	-11	08	02	08	-07	03	13	08	-15	-17	-22	-06	19	-17	-01	-26	-04	01
9.									-25	09	-15	-07	09	-16	-02	13	10	-03	-08	-02	01	-13	-01	-09	-06	02	-14	-03
10.										-10	45	-14	04	28	07	02	-03	13	07	-17	-03	24	-01	20	-01	01	08	-03
11.											-07	23	09	12	-17	19	10	16	-25	-04	-18	03	33	-32	12	-14	-16	13
12.												-02	04	16	04	-01	-05	13	11	-04	09	25	-03	-09	04	20	15	00
13.													18	-01	-01	-05	-07	09	-01	-03	-07	-10	-25	-16	11	-06	05	-06
14.														11	12	-04	-17	17	11	11	07	07	21	-00	15	13	-04	11
15.															22	18	-28	-01	30	20	16	26	-01	15	22	14	25	15
16.																-27	-22	-09	40	14	11	-14	-17	20	04	21	11	11
17.																	38	19	-18	-04	-04	-00	09	-09	02	-39	-18	-09
18.																		14	-23	01	-07	-00	09	-14	-03	-35	-18	-04
19.																			03	-22	-02	14	10	-06	-03	-26	-05	-16
20.																				20	22	00	-11	16	06	24	50	21
21.																					13	06	03	08	07	09	15	21
22.																						30	-11	00	02	14	17	09
23.																							-03	-02	05	-00	-01	03
24.																								-23	12	-12	-07	09
25.																									02	18	08	-04
26.																										09	08	03
27.																											15	21
28.																												19

* Decimal points are omitted; numbers referring to statements correspond to the numbers accompanying the statements in Table 1.

TABLE 3

R-FACTOR SIMPLE STRUCTURE MATRIX N = 130, VARIMAX ROTATION*

Statement	Factor 1	Factor 2	Factor 3	Factor 4	h²
1	.690	-.005	.240	.064	.539
2	.146	-.476	-.090	.246	.317
3	-.173	.065	-.020	.522	.306
4	.707	-.085	.016	-.257	.573
5	.654	.061	-.153	-.163	.481
6	-.034	-.064	.102	.615	.393
7	.648	.128	.065	-.102	.450
8	-.196	.153	.168	.174	.120
9	.028	-.388	.014	.214	.197
10	.165	.650	-.105	.151	.483
11	-.273	-.025	.510	.198	.375
12	.293	.507	-.021	.099	.353
13	-.030	-.077	.400	.056	.170
14	.137	.022	.381	.026	.165
15	.461	.298	.305	-.082	.401
16	.539	-.124	.037	-.119	.321
17	-.217	.015	.041	.476	.276
18	-.377	-.019	-.092	.331	.261
19	-.088	.317	.073	.224	.164
20	.588	.066	.094	-.353	.483
21	.125	-.156	.189	-.144	.096
22	.251	.074	-.026	-.161	.095
23	.006	.451	.012	-.003	.203
24	-.401	.097	.488	-.051	.411
25	.394	-.009	-.266	.033	.227
26	.113	.082	.359	.002	.148
27	.450	-.090	.025	-.363	.343
28	.325	.205	.125	-.330	.272
29	.140	-.061	.282	-.159	.128
Percent Variance	15.66	5.79	4.66	4.07	30.18

* Principal axis solution; minimum eigenvalue criterion = 1.000. Four chosen eigenvalues = 4.5402, 1.6788, 1.3528, 1.1805.

Traditionally Negro institutions serve a necessary function by offering the Negro student a curriculum which more nearly meets his needs and educational background.

All respondents also disagreed with the statement that "running a university is basically like running a business." Respondents disagreed mildly with statements that collective bargaining should not be used by a university faculty, and that more federal money for higher education will bring more federal control.

Four consensus statements with which there was agreement revolved around making higher education available to anyone who wants it, giving special considerations to disadvantaged students, designing curricula to serve highly divergent needs and interests, and encouraging active participation by the university in solving social problems. Opinions about needed changes in the grading system were mixed, although the differences were not substantial.

Correlations between z-score patterns for the three opinion types appear in Table 4. Although two of the coefficients were statistically significant, the relationships were low, the largest accounting for only 20.25 percent of the variance. The z-score patterns for all opinion types appear in Table 1.

The most prominent characteristic of Type I (N = 52) seemed to be a concern for academic and personal freedom of both faculty and students and a concern for faculty participation in the governing of the university. To a greater extent than all other types, Type I felt a need for more faculty representation on the university's governing board.

TABLE 4

CORRELATIONS BETWEEN z-SCORE OPINION TYPES

Opinion Type	II	III
I	.28	.45*
II		.43*

* $p < .05$

Type I also felt faculty members should not be subject to administrative reprisal for expressing their opinions freely, and that the faculty should have access to university communication channels (the student newspaper) to make its thoughts known. Type III agreed on both counts, albeit to a substantially lesser extent, while Type II disagreed on the latter point. Type I also would extend freedom to speak to others as evidenced by its strong disagreement with this statement:

All campus speakers should be subject to some official screening process.

Types II and III felt all campus speakers should be screened before being allowed to speak.

Type I disagreed only slightly with the statement that students who disrupt the campus should be expelled or suspended; other types felt such students should be expelled. Type I also strongly disagreed with the idea that the administration should exercise control over the content of the student newspaper; Type III disagreed only mildly, and Type II favored administrative control.

To a lesser extent than did Type II, Type I disagreed with the statement that the PhD and EdD are overemphasized in recruiting new faculty; Type III felt such an overemphasis exists. Type I also, less than the other types, disagreed with the statement that the university president should serve more as a mediator than a leader. Finally, while the other types disagreed, Type I agreed with this statement:

An active research interest is a prerequisite for good undergraduate teaching. A man who does no research on a subject soon becomes less qualified to teach it.

What are the demographic characteristics of Type I? Analysis of data in the multiple linear regression accounted for 32.48 percent of the variance of the Type I factor loadings (see Table 5). Three variables accounted for significant proportions of variance, and the Type I individual is most likely to be a "liberal" and a Democrat who has been on the campus less than 10 years.

Type II (N = 27) displayed what might be termed a power and control orientation, perhaps an academic version of the current political "law and order" slogan. Typifying this characteristic was the following statement, with which Type II agreed and the other types disagreed:

The (name of university) administration should exercise control over the content of the student newspaper.

Perhaps more significant was Type II's noncommittal response to the statement that it is reasonable to require faculty to sign a loyalty oath. Both Types I and III felt strongly that the requirement is unreasonable. Type II did not believe faculty should have the right to express their opinion freely on any issue through university communication channels; while Types I and III strongly objected to college officials disciplining students who take part in civil disobedience off campus, Type II felt only slightly that the administration should not take punitive action. Type II strongly disagreed with the statement that the typical undergraduate curriculum has "suffered from the specialization of the faculty" and disagreed very strongly with the statement that the importance of the PhD and EdD is overemphasized in recruiting new faculty. Type I disagreed on both of these points to a substantially lesser degree than did Type II, and Type III agreed with both points. Type II also strongly felt that attending the university is a privilege, not a right, and that the university should be as concerned with the personal values of students as with their intellectual development.

Two demographic variables—political party

TABLE 5

PREDICTOR VARIABLES ACCOUNTING FOR SIGNIFICANT PROPORTIONS OF VARIANCE FOR EACH OPINION TYPE

Variable	Variance	p-value	r_{pc}*
Type I			
Liberal	9.48	< .01	.4835
Democrat	18.67	< .01	.4516
Less than 10 years on campus	4.33	< .05	.2105
TOTAL	32.48		
Type II			
Liberal	4.55	< .025	-.3566
Republican	16.26	< .01	.4033
TOTAL	20.81		
Type III			
Liberal	3.25	< .05	-.2256
Research	6.65	< .01	-.2506
More than 10 years on campus	5.78	< .05	.2404
TOTAL	15.68		

* zero order correlation coefficient between predictor and criterion variable.

preference and political orientation—accounted for 20.81 percent of the factor loading variance. Type II is most likely to be a Republican, and he is quite unlikely to consider himself a "liberal."

Differentiating Type III (N = 30) from the other types was its "teaching versus research" orientation. Type III most strongly agreed with this statement:

> Teaching effectiveness, not publications, should be the primary criterion for promotion of faculty at (name of university).

Types I and II also agreed with the statement but to a substantially lesser degree. Type III strongly disagreed with the statement that "an active research interest is necessary for good undergraduate teaching" and that a faculty member who does no research "becomes less qualified" to teach a subject. Type II also disagreed but to a substantially lesser extent, and Type I agreed. Type III felt strongly that the value of the PhD and EdD is overemphasized in recruiting new faculty. Type III did not feel— as did the other types—that in making admissions decisions, academic aptitude should be the most important criterion.

Three demographic characteristics accounted for

15.68 percent of the variance of the Type III factor loadings. Type III has been on campus for more than 10 years and almost certainly is not involved in research. He does not consider himself a "liberal."

DISCUSSION

Any attempt to identify the concerns and opinion patterns of a university faculty involves certain encumbrances difficult to dislodge. Obviously, results pertain only to the particular time of the investigation and, when involving a single institution as did this study, to a particular university. In addition, this study dealt with a limited set of statements as well as a limited set of respondents. Thus, while the three opinion patterns may be representative of the feelings of a large proportion of faculty on campus, they cannot be considered as reflecting either the "average" faculty member's opinion or all of the possible opinions held by the faculty on these issues. Other sets of statements with these and other faculty members may well produce somewhat different responses.

Nonetheless, it may be presumed that these three opinion types do reflect some of the diversity of opinion that exists within a university faculty. Results certainly lend empirical support to speculation about

the division of thinking among today's faculty. Rather sharply delineated, for example, was the split between teaching and research orientations. Similarly apparent was the division between those calling for more administrative control and those desiring less such control. Results also pointed up strong faculty concern for the university functioning as a social catalyst.

Efforts to describe demographic characteristics of opinion Types II and III were not particulary successful, as evinced by the relatively small proportions of variance accounted for by regression on demographic variables. Despite this, Q-factor analysis and multiple linear regression would appear to be useful methods in achieving better descriptions of opinion patterns relevant to issues of faculty concern and in identifying individuals reflecting differing opinion patterns. The need is for replication and more extensive study. Such efforts could help in understanding the position of the faculty member in today's rapidly changing structure of higher education.

FOOTNOTES

1. The authors gratefully acknowledge cooperation of the Education Testing Service and, particularly, ETS's permission to adapt questions from the College Trustee Study.

2. Factor and z-score matrices for the Q-analysis are available from the authors upon request. Send requests to Dr. Kenneth Starck, Department of Journalism, Southern Illinois University, Carbondale, Illinois 62901.

REFERENCES

1. Cattell, Raymond B., *Handbook of Multivariate Experimental Psychology*, Rand McNally and Company, Chicago, Illinois, 1966.

2. Dressel, Paul L.; Lorimer, Margaret F., *Attitudes of Liberal Arts Faculty Members Toward Liberal and Professional Education*, Institute of Higher Education, Teachers College, Columbia University, New York, New York, 1960.

3. Furst, Edward J., "A Factor Analysis of Preferences in Teacher Role-Behavior," *The Journal of Experimental Education*, 33:379-382, Summer 1965.

4. Graybeal, William S., "What the College Faculty Thinks," *NEA Journal*, 55:48-49, April 1966.

5. Harman, Harry H., *Modern Factor Analysis*, *Second Edition Revised*, University of Chicago Press, Chicago, Illinois, 1967.

6. Kelly, Francis J. and others, *Research Design in the Behavioral Sciences: Multiple Regression Approach*, Southern Illinois University, Carbondale, Illinois, 1969.

7. Lorimer, Margaret F.; Dressel, Paul L., "Faculty Characteristics—College and University," in Ebel, Robert L. (ed.), *Encyclopedia of Educational Research, Fourth Edition*, The Macmillan Company, New York, 1969.

8. "Most College Trustees Found White, Protestant, Republican," *The Chronicle of Higher Education*, January 13, 1969.

9. Richardson, Richard C., Jr.; Blocker, Clyde E., "An Item Factorization of the Faculty Attitude Survey," *The Journal of Experimental Education*, 34:89-93, Summer 1966.

10. Russell, John Dale, "Faculty Satisfactions and Dissatisfactions," *The Journal of Experimental Education*, 31:135-139, Winter 1962.

11. Stephenson, William, *The Study of Behavior*, The University of Chicago Press, Chicago, Illinois, 1953.

DEVELOPMENT OF A SEMANTIC DIFFERENTIAL TO ASSESS THE ATTITUDE OF SECONDARY SCHOOL AND COLLEGE STUDENTS

RUSSELL N. CASSEL
University of Wisconsin-Milwaukee

ABSTRACT

The inquiry sought to develop a semantic differential (SD) for use in assessing attitude and attitude change among secondary school and college students. It included thirty-five bipolar adjectives, each using a 7-point ordinal scale, for student rating purposes. Three concepts were used in the study: teacher, learning, and student. A Likert type scoring was accomplished with part scores for each of the separate concepts, and with the total score being the sum of the three part scores. A comparison between pre- and post-college course attitudes was made involving 237 students, which showed significant change only for the concept "student" (Me as a student). Internal reliability indexes were obtained using the Kuder-Richardson(K-R) Formula 20 for part scores ranging from r = .421 to .610; and for the total score ranging from r = .928 to .960. Intercorrelations of part scores for pretest ranged from r = .530 to .584; and for posttest r = .620 to .707. There is evidence of greater homogeneity for post-course concepts than for pre-course concepts used in the evaluation, i.e., teacher, learning, and student.

THE OBJECTIVE of the inquiry was to develop a psychological instrument for use in assessing the attitude of secondary school and college students. It sought to establish semantic scales for use in a semantic differential on the basis of rigid adherence to usual test development and standardization procedures (5), to use adjectives for the development of semantic scales that proved to be critical in previous studies, to use Likert type scoring of the semantic differential with part scores for separate "concepts," and to validate against meaningful criterion variables.

DEVELOPMENT OF SEMANTIC DIFFERENTIAL

Inasmuch as the reliability of a psychological instrument is in large part a function of the number of items contained as a "sample of behavior" of criteria being assessed, it was deemed that a minimum of thirty or more semantic scales, used as individual items, would be essential.

DEVELOPMENT OF SEMANTIC SCALES

Each semantic scale was comprised of a rating scale anchored by bipolar adjectives, and as traditionally used for the semantic differential (6).

The Adjective Check List, by Gough and Heilbrun (2), has been used extensively in connection with the identification and evaluation of adjectives for their criticalness in relation to human behavior. Accordingly, twenty of the thirty-five semantic scales used in the final standardized Semantic Differential for Secondary Students (see Appendix) were developed from adjectives suggested in studies as being critical by studies using The Adjective Check List. Ten of these adjectives were reported by Applezweig (1) in a study involving 360 entering students at Connecticut College for Women. Five of these adjectives were selected by freshmen women with superior grades at the end of the first semester, all of whom were on the Dean's list, and with opposite adjectives as follows:

practical - imaginative
thorough - partial
logical - illogical
sympathetic - unsympathetic
appreciative - unappreciative

The other five adjectives were selected by freshmen women with inferior grades at the end of the first semester, all of whom were on probation, and with opposite adjectives as follows:

affectionate - hateful
forgiving - unforgiving
frank - deceitful
loyal - disloyal
tolerant - intolerant

Ten more of the adjectives were reported in The Adjective Check List Manual (2) for a study of 295 males, with six coming from those having high scores on the "Mathematician Scale" of Strong Vocational Interest Blank, and four from those having low scores:

High Scores
civilized - uncivilized
curious - indifferent
insightful - blind
original - imitational
rational - irrational
sensitive - insensitive

Low Scores
lazy - ambitious
narrow interest - broad interest
shallow-deep
simple - complex

The remaining 15 semantic scales were taken from studies that clearly indicated the factorial identity of each (4, 5, 6):

Evaluative Factor
wise - foolish
successful - unsuccessful
valuable - worthless
honest - dishonest
interesting - boring
pessimistic - optimistic

Familiarity Factor
clear - vague
usual - unusual
disorderly - orderly
conservative - progressive

Activity Factor
active - passive
excitable - calm
inhibited - uninhibited

Potency Factor
strong - weak
fast - slow

SCALES USED FOR RATING PURPOSES

A 7-position ordinal scale was interposed between each pair of bipolar adjectives forming the thirty-five semantic scales. The seven positions on each scale were defined as follows: (1) extremely, (2) moderately, (3) slightly, (4) neutral, (5) slightly, (6) moderately, and (7) extremely. The subject is asked to rate a concept on the 7-point scale in terms of which of the two bipolar adjectives is believed to be most appropriate, and in terms of

the four adjective positions adjacent to such word.

Concepts Used

Three different concepts were used in the standardization of The Semantic Differential for Secondary School Students (S-D):

What I learned in this class.
The teacher of this class.
Me as a student.

Each one of the three different concepts made use of the same thirty-five semantic scales described.

STANDARDIZATION

Six hundred and ten student records were used in the standardization process. About half of them 287, were from high school students; while the remainder, 323, were from upper-division college students or graduate students.

Item Retention and Revision

All semantic scales were subjected to an item analysis, and only those items that correlated .20 or better with the total score for all three concepts were retained. Three separate revisions were necessary before reasonable stability was established, and where an r of .20 or better was established for two of the three concepts utilized, i. e., (1) What I learned in this class, (2) The teacher of this class, and (3) Me as a student.

Assigning Weights to Semantic Scales

Each of the thirty-five semantic scales was assigned values ranging from 1 to 7 for the seven adjective positions on the interposed ordinal scales. The initial step in the weighting involved identifying those adjective pairs where one of the adjectives seemed clearly to be desired to the other, and the value of 7 was assigned to the side of the semantic scale with that adjective, with the 1 being assigned to the other, i. e., practical, thorough, logical, appreciative, honest, loyal, and the like. A statistical technique was then used to determine on which side the value of 7 was to be assigned on the semantic scales where it seemed questionable which adjective of the bipolar pairs was to be desired, i. e., original, active, excitable, narrow interest, conservative, fast, tolerant, etc. (7).

Reliability

Data contained in Table 1 illustrates internal consistency type of reliability for each of the three part and the total scores by use of the traditional K-R Formula 20. The part scores range from an r of .421 for Part II - Learning, to an r of .610 for Part III - Student. Total score reliabilities were computed for three different variations of the K-R 20 Formula, i. e., Traditional K-R 20 assumes all items have equal difficulty and correlations; Cronbach Alpha obtains correlation for all possible splits of the test; while Horst corrects for dispersion of item difficulty. When there is little dispersion of item difficulty, there is little difference among the r's obtained for the three variations of the K-R Formula

TABLE 1

INTERNAL CONSISTENCY RELIABILITY (N=610)

Variation of K-R 20	Part I Teacher	Part II Learning	Part III Student	Total S-D Score
Traditional	.534	.421	.610	.928
Cronbach's				.929
Horst's				.960

20. Since the r obtained for the Horst variation of the K-R Formula 20 is considerably larger than for the traditional K-R 20 and the Cronbach Alpha, it is obvious that there was considerable dispersion of item difficulty in the study.

Scoring of S-D

Three part scores were computed on the Likert technique (3). Each semantic scale (test item) received a weight from 1 to 7, with the value being assigned to the right side when a single asterisk follows the scale, and on the left when two asterisks follow the scale as shown in the appendix. The same thirty-five semantic scales were used for all three concepts, with each concept representing a part score, and with the sum of the three part scores being the total score on the S-D:

> Part I - Teacher: (rating for "The teacher of this class"),
> Part II - Learning: (rating for "What I learned in this class"),
> Part III - Student: (rating for "Me as a student in this class"), and
> Total S-D Score: sum of the three separate scores.

TABLE 2

INTERCORRELATIONS OF PRE-COURSE S-D SCORES (N = 243)

Scores on S-D	Part I Teacher	Part II Learning	Part III Student	Total S-D Score
Part I - Teacher		.538	.584	.876
Part II - Learning			.530	.807
Part III - Student				.908
Total S-D Score				
Mean	191.81	173.76	166.69	532.25
Standard Deviation	24.73	28.96	18.88	63.97

Intercorrelations of S-D Scores

The intercorrelations of scores on the S-D were computed separately for pre- and post-course administration and as illustrated in Tables 2 and 3 respectively. The means and standard deviations for the pre- and post-course administration of the S-D were also included. The shift in change of correlation coefficients is in a direction of greater common variance between the student and both the teacher and learning, and with the greatest shift being towards embracing values of teacher, i.e., from an r of .584 to an r of .707. By comparing the means for the pre-course S-D administration from Table 2, with the post-course means in Table 3, it can be seen that the greatest change takes place with student, as opposed to the teacher and learning.

Criterion Study

Two faculty members from the Educational Psychology Department at the University of Wisconsin-Milwaukee, an assistant professor and a full professor, were involved. Two hundred and forty-three students were asked to score the S-D at the beginning of the semester, but only 237 of the same students completed it at the end. An analysis of variance for correlations observations was accomplished to determine if there was a significant change in the attitude of students as measured by the S-D for the three concepts included. The data illustrating the findings of that test are contained in Table 4. The only statistically significant change indicated in Table 4 is for Part Score III - Student, and which deals with the student's own opinion of himself as a student. The change is in a direction of greater esteem for self, with little or no significant change in either the teacher or what he thought he learned during the particular course involved. Based on this finding, students appear to feel they have changed for the better as a result of the courses, but the basis of that change does not appear to involve a change in attitude toward either the teacher or what they have learned.

Factor Analysis

Two separate principal component factor analyses were accomplished involving the 105 semantic scales (items on the S-D) as variables, and the entire 610 subjects involved in the initial standardization process. The data for these two analyses are not included, as they are too voluminous, and contribute little to the findings. The first of the two analyses extracted twelve separate factors with one eigenvalue or more, with the first factor accounting for 60 percent of the total variance, and all twelve factors accounting for 83 percent. When these twelve factors were rotated to simple structure by use of the varimax orthogonal method, four factors under each of the three concepts were obviously in agreement with the Osgood factor content of semantic scale initially included: I - Evaluative, II - Activity, III - Familiarity, and IV - Potency.

The second factor analysis was done with the same data, and in the same manner, except that only three factors were extracted. This was done to determine if the factorial content for the three concepts (teacher, learning, and student) was more potent than the factor identification of the semantic scales. The

TABLE 3

INTERCORRELATIONS OF POST - COURSE
S - D SCORES (N = 237)

Scores on S - D	Part I Teacher	Part II Learning	Part III Student	Total S - D Score
Part I - Teacher		.687	.707	.843
Part II - Learning			.620	.812
Part III - Student				.878
Total S - D Score				
Mean	194.80	174.84	187.38	557.02
Standard Deviation	21.53	28.09	20.73	62.33

first of the three factors accounted for 60 percent of the total variance, but all three of the factors only accounted for 70 percent of the variance. Interaction of the semantic scales and concepts seem to follow the pattern described by Nunnally (5) and others, where the loadings for the concepts are more factorially potent than for the semantic scales.

REFERENCES

1. Applezweig, D. G., "Self-Perception in Later Adolescence," paper read at International Congress of Psychology, Bonn, Germany, 1960.

2. Gough, H. G., Heilbrun, A. B., The Adjective Check List Manual, Consulting Psychological Press, Palo Alto, California, 1965.

3. Likert, R. "A technique for the Measurement of Attitude," Archives of Psychology, Number 140, 1932.

4. McNeil, K. A., "Semantic Space as an Indicator of Socialization," Journal of Educational Psychology, 59: (no. 5) 325-327, 1968.

5. Nunnally, J., Psychometric Theory, McGraw Hill, New York, 1967.

6. Osgood, C. E.; Suci, G. J.; Tannenbaum, P. H., The Measurement of Meaning, University of Illinois Press, Urbana, 1957.

7. Torgerson, W. S., Theory and Methods of Scaling, John Wiley and Son, New York, 1962.

APPENDIX

THE SEMANTIC DIFFERENTIAL FOR SECONDARY SCHOOL STUDENTS

This SEMANTIC DIFFERENTIAL is intended for use in assessing the attitude of persons in relation to certain concepts related to learning. It consists of thirty-five Semantic Scales (adjective antonyms) which are related to effectiveness in student learning; both at the secondary and college levels of instruction. The three important concepts which have been used in the preliminary validation of this instrument are: (1) Teacher, (2) Learning, and (3) Student. Any number of other pertinent concepts may be used with the same thirty-five Semantic Scales contained in this instrument.

General Directions:

Each of the following pages contains a 'concept' at the top which is believed to be related to how well you have learned, and with thirty-five different pairs of opposite adjectives, which are called 'Semantic Scales.' The concept is different for each page, but the thirty-five adjective pairs are the same. You are to mark each of the thirty-five adjective pairs, which we will refer to as 'Semantic Scales,' in relation to how you actually feel about the concept at the top of the page. The concept at the top of the first page is " WHAT I LEARNED IN THIS CLASS, " and we have used this concept in the Example that follows:

EXAMPLE: WHAT I LEARNED IN THIS CLASS

EX   MO   SL   NE   SL   MO   EX*

(1) Strange X : ___ : ___ : ___ : ___ : ___ : ___ Familiar

(2) Ugly ___ : ___ : ___ : ___ : ___ : ___ : X Beautiful

(3) Easy ___ : ___ : ___ : X : ___ : ___ : ___ Hard

* EX = Extremely: MO = Moderately: SL = Slightly: NE = Neutral: SL = Slightly: MO = Moderately: EX = Extremely:

If you think that " HOW WELL YOU LEARNED IN THIS CLASS " was strange, make an " X " near the word strange; but if you think it was more familiar, mark the " X " near the word familiar. The example with the " X " right next to strange indicates that the student thought WHAT HE LEARNED IN THIS CLASS was strange. The " X " right next to beautiful suggests that he felt WHAT HE LEARNED IN THIS CLASS was beautiful. For the " Easy-Hard " Semantic Scale (adjective antonyms) the " X " is placed right in the middle of the scale, or about half-way between easy and hard. This is a neutral position indicating that the student felt that WHAT HE LEARNED IN THIS CLASS was neither easy nor hard. It was probably some of each; so he placed the " X " in the middle of the Semantic Scale for that concept.

Remember each page contains a new and different concept, but the same thirty-five Semantic Scales are used. You are to mark each of the thirty-five Semantic Scales for all of the concepts included. When you are finished, turn the booklet face-down.

WHAT I LEARNED IN THIS CLASS

	Extre-mely	Moder-ately	Sligh-tly	Neut-ral	Sligh-tly	Moder-ately	Extre-mely	
1. practical		:	:	:	:	:	:	Imaginative*
2. thorough		:	:	:	:	:	:	partial**
3. logical		:	:	:	:	:	:	illogical**
4. sympathetic		:	:	:	:	:	:	unsympathetic**
5. clear		:	:	:	:	:	:	vague**
6. appreciative		:	:	:	:	:	:	unappreciative**
7. civilized		:	:	:	:	:	:	uncivilized**
8. curious		:	:	:	:	:	:	indifferent**
9. insightful		:	:	:	:	:	:	blind**
10. original		:	:	:	:	:	:	imitational**
11. rational		:	:	:	:	:	:	irrational**
12. sensitive		:	:	:	:	:	:	insensitive**
13. wise		:	:	:	:	:	:	foolish**
14. interesting		:	:	:	:	:	:	boring**
15. successful		:	:	:	:	:	:	unsuccessful**
16. strong		:	:	:	:	:	:	weak**
17. active		:	:	:	:	:	:	passive**
18. fast		:	:	:	:	:	:	slow**
19. usual		:	:	:	:	:	:	unusual*
20. valuable		:	:	:	:	:	:	worthless**
21. excitable		:	:	:	:	:	:	calm**
22. honest		:	:	:	:	:	:	dishonest**
23. affectionate		:	:	:	:	:	:	hateful**
24. forgiving		:	:	:	:	:	:	unforgiving**
25. frank		:	:	:	:	:	:	deceitful**
26. loyal		:	:	:	:	:	:	disloyal**
27. tolerant		:	:	:	:	:	:	intolerant**
28. pessimistic		:	:	:	:	:	:	optimistic*
29. lazy		:	:	:	:	:	:	ambitious*
30. narrow interest		:	:	:	:	:	:	broad interest*
31. shallow		:	:	:	:	:	:	deep*
32. simple		:	:	:	:	:	:	complex*
33. conservative		:	:	:	:	:	:	progressive*
34. inhibited		:	:	:	:	:	:	uninhibited**
35. disorderly		:	:	:	:	:	:	orderly*

EFFECTS OF NEUROLOGICAL TRAINING

ON PSYCHOMOTOR ABILITIES

OF KINDERGARTEN CHILDREN

RICHARD D. CORNISH
Unified School District Number 1, Racine, Wisconsin

ABSTRACT

That neurological dysfunction accounts for many cases of academic failure is well documented, but specific research on programs designed to overcome this dysfunction is sparse. In this study fifty kindergarten children having perceptuomotor and/or psychomotor deficits as identified by several well-known scales were assigned to an experimental or control group. The experimental group was then given a program of cross-patterning exercises for 3 minutes a day over a 3-month period. The obtained results suggest that a neurological training program does not significantly improve the psychomotor functioning of kindergarten children.

THE CAUSES of academic failure (children who benefit to a limited degree from instruction received) are many. Indeed, there may be as many causes as there are academic failures. Research into the etiology of failure has been carried out by educators, psychologists, psychiatrists, pediatricians, and sociologists to name a few. Causal factors delineated by these researchers by and large fall into their own area of specialization, but basically into five categories: (a) congenital defect or deficit, (b) environmental influence, (c) psychological factors, (d) physiological factors, or (e) various combinations of the preceding four (16).

In short, research has not delineated a single causal factor of academic failure. McLeod (16), while recognizing that etiological factors must eventually be treated, holds that they are so diverse, and methods of identifying them so unsatisfactory, that for now, the symptomatic behavior is all that can be treated, i. e., remedial reading programs, etc. However, problem children typically cannot be identified until long after they should have started to read—the end of the second grade at the earliest. This forces all programs to be remedial and leaves no room for prevention that would be possible if academic failure could be recognized earlier, or could be predicted.

This is no small problem. Estimates of the numbers of children whose academic achievement does not correspond with intellectual potential run as high as 30 percent of the school district population (4). Although McLeod (16) holds that etiological factors are too varied or evasive to be of concern, he does recognize that neurological dysfunction accounts for many causes. He states, "A study of the behavior of the undeveloped learner will reveal the majority of the children experiencing difficulties in developing their basic learning skills also have difficulty organizing their visual, auditory, and motor experiences (16:27).

For reasons of parsimony, all types of academic failure will not be discussed here. In the ensuing discussion, reading disability will be assumed to be concomitant with and indicative of basic academic failure. There is ample evidence to support the fact that visual, auditory, and motor learning skills are fundamental to learning to read successfully, the same factors found in general disability (3, 9, 10, 19, 21).

That reading failure can be predicted has been shown by DeHirsch, Jansky, and Langford (5). In speaking of children who are subsequently poor readers, these authors note that as early as kindergarten, "They show difficulties with fine motor control, a crude body image, and primitive visuomotor

experiences, all testifying to . . . a generalized developmental dysfunction" (5:xiii). A battery of tests were developed to reflect perceptuomotor and linguistic ability at the kindergarten level. This battery subsequently proved to have predictive validity for reading difficulties.

That reading difficulty can be predicted is not a new idea. It had been advocated by Delacato (6) as early as 1959. Delacato's work, however, contained no empirical verification on the prediction of reading problems. He held, through clinical judgment, that reading problems could be prevented. Delacato (6) also presented a method for this prevention, a method that has become synonomous with his name. The method is long and involved, encompassing such things as handedness, eye preference, thumb sucking, posturalization, etc. The Delacato theory raised much controversy and, because of its complexity, generated very little empirical research.

In 1963 Delacato (7) redefined his position and more clearly delineated his treatment and prevention procedures. He claims his new work was the product of much research, but fails to report this research as research. The reader often cannot separate research findings from clinical intuition. In the 1963 work the major technique in the treatment and prevention schema is that of cross-patterning, a technique that has received subsequent research attention and the one to which the present study directs itself.

The present study grew out of the author's work with the Oconomowoc, Wisconsin, Public School's Learning Center. During the summer of 1968, 372 children were enrolled in the Center's Pre-Kindergarten Clinic. Of this number 107, or just over 29 percent, were found to be retarded or deficient in basic perceptuomotor skills. During the following year, while the children were in kindergarten, a cross-patterning program was instituted to see if the perceptuomotor difficulties could be overcome. The specific hypothesis tested in this study is that children with perceptuomotor difficulties who are given cross-patterning exercises will show a greater improvement in perceptuomotor coordination than children with the same difficulties who are not given the exercises. It is further hypothesized that the above differences will be significant at the .05 level.

RELATED RESEARCH REVIEW

There has been very little research done in this area. Most of what is available is found reported by Delacato (8) in ". . . ten carefully controlled experiments. . . ." However, these studies have many weaknesses.

Sister Mariam (12) compared reading disability of neurologically organized and neurologically disorganized children in a descriptive study. No attempt was made to improve neurological organization or reading ability. The study showed there is a significant difference in reading ability between neurologically organized and neurologically disorganized children. The neurologically disorganized children were found to score significantly lower in the areas of comprehension, visual and auditory recognition, and oral reading performance.

Masterman (13) dealt with neurological training exercises and their effect on reading retardation.

One hundred forty-one pairs of children were matched according to sex, age, and grade placement. One member of each pair was in the experimental group and received various neurological training exercises, including cross-patterning. The control group received no treatment. Mean gain scores on the Gray Oral Reading Test were compared and the probability that both groups came from the same population was found to be "near the .01 level."

The Masterman study contains several weaknesses. Firstly, Masterman stated that any differences found would be in favor of the experimental group and hence used a one-tailed test. Although this is a possible assumption, the differences may not necessarily be in one direction and, as such, a two-tailed test should have been used.

Secondly, the treatment period lasted for 1 month. In that the scores on the Gray Test are based on a 10-month year, Masterman multiplied the pre-posttest differences scores by 10 for the purpose of analysis. This is based upon the assumption that learning curves are increasing linear functions.

McGrath (15) found that ninety-two students, ranging from grade 7 to grade 11, enrolled in a summer remedial reading program, read at fifth to ninth grade levels. All ninety-two students were given the Metropolitan Reading Test, form Am as a pretest. All students were then given a remedial reading program, consisting of the Science Research Associates Reading Kit, the Reader's Digest Practice Reader, and the Catholic Charities Spelling Charts, in addition to a neurological training program, including cross-patterning. All students were given five 5-day weeks of this program and then were retested on the Metropolitan Reading Test, form Bm. McGrath reported his results verbally, however Delacato supplied a data analysis. He subtracted the expected difference in pre- and posttest scores from the obtained difference and divided the answer by the standard error of the mean difference and found significant improvement at the .01 level. However, without appropriate control groups, the obtained difference cannot be attributed to any of the treatments.

Kabot (11) studied experimental and control groups consisting of eleven matched pairs of third-grade children who, on the Stanford Reading Achievement Test, scored 6 months or more lower than their age level. Following 8 weeks of neurological training, the two groups were given a posttest, using the California Reading Test. The results obtained when comparing the improvement of the experimental and control groups were not significant.

Correlations between the Stanford Test and California Test are not reported. In this case, the training program may have been effective and this difference lost due to the use of two completely different tests.

To assess any possible transfer of training, both of the experimental and control groups were tested on an alternate form of the California Test, after receiving a 7-month remedial reading program. The resulting difference between the experimental and control groups was taken as evidence that neurological training transferred and resulted in significantly improved performance in a remedial reading program.

Alcuin (1) divided 120 Ss (method not given) into

three groups; the Reading, Psychological, and Neurological Groups. Each group was given their respective treatment three times a day, twenty times per session over a period of 6 weeks. During these three periods the Reading Group did work related to reading, e. g., penmanship. The Neurological Group engaged in a neurological training program including cross-patterning exercises. The Psychological Group did calisthenics, "so planned as not to involve the whole organism in such a way as to improve Neurological Organization" (1:152). All groups were given the Stanford Reading Test, both before and after the treatment period. An analysis of variance on the mean gain scores showed a significant difference between the groups. Subsequent t-tests showed the Neurological Group to have improved significantly more than the Reading or Psychological Groups, and that the two latter mentioned groups did not differ significantly from each other.

Although this area is starting to receive attention in the professional journals (14, 17), research material other than that reported by Delacato (8) is yet sparse. A study by Painter (18) was designed to "investigate the effects of a rhythmic and sensory motor activity program on body image, perceptual motor integration, and psychological competence of kindergarten children."

The Ss for the experiment were the lower 50 percent of a forty student kindergarten class. Relative standing in class was indicated by Goodenough MA scores. The twenty Ss were then divided into matched pairs on the basis of chronological age, (CA), mental age (MA) and sex.

The experimental group was given training sessions three times a week over a 7-week period. No time was spent with the control group to compensate for the Hawthorne effect. The treatment method used was not that advocated by Delacato (6, 7), but one patterned after that suggested by Barsch (2). Many of these exercises bear a strong functional relationship to those of Delacato, including jumping, hopping, skipping, and bilateral body movement.

In this study the author did not report results but merely listed the following hypotheses and reported levels of significance: the treatment procedures would affect the level of performance in drawing a human figure: $p < .055$, sign test; the program will correct distortion of the body image concept: $p < .01$, sign test. The treatment will improve motor integrity: $p = .004$, sign test; the program will improve sensory motor spatial performance skills: $p = .002$, sign test; the procedure will improve psycholinguistic abilities: $p = .055$, sign test.

Noting that the Delacato method is so long and involved that controlled experiments are difficult and, when done, cannot be replicated, Silver, Hagen, and Hersh (20) attack the problem of reading disability through the direct stimulation of the deficit perceptual areas. Ss for the experiment were eighty males from 7 to 11 years old, all of whom had been referred to a mental hygiene clinic for learning and behavior problems. The Ss were paired on the basis of age, IQ, and neurological and psychiatric examination, then randomly assigned to one of the two experimental groups.

The training sessions were individual, each lasting 45 minutes and held twice weekly over a period of 1 year. For the first 6 months one group received perceptual stimulation while the other group received "contact appointment"—conventional teaching of reading by a tutor. For the second 6 months the situation was reversed. Hence, E_1 and E_2 differ only in their order of presentation of the treatments. All children were given the criterion test battery at the beginning of the experiment, at the end of 6 months, and at the end of the experiment. Thus, in addition to the group contrasts, each S acts as his own control. The training techniques are not delineated, but were stated to cover the visual, auditory, tactile, and kinesthetic modalities as well as body image.

The experiment was not completed at the time of publication, so no quantitative data was presented. Two case studies were presented and the preliminary results "showed definite progress."

METHOD

Subjects

The Ss were fifty kindergarten children who had been found to have perceptuomotor and/or psychomotor deficits, as defined by scores on subtests of the Purdue Perceptual Motor Survey (walking board, ocular pursuit, identification of body parts, imitation of movement), the Draw A Man and Ten Dot subtests of the Anton Brenner Developmental Gestalt Test of School Readiness, and lack of visual fusion as measured by the Keystone Visual Survey. The Ss were assigned to the experimental (E) or control (C) group, depending upon the school they attended. Schools were randomly assigned to treatment. E contained twenty-three Ss, nine males and fourteen females. C contained twenty-seven Ss, sixteen males and eleven females.

Tests

Ss were both pre- and posttested on the Purdue Perceptual Motor Survey, the Keystone Visual Survey, and the Draw A Man and Ten Dot subtests of the Anton Brenner Developmental Gestalt Test of School Readiness.

Apparatus

The apparatus used in this experiment was the Exer-Cor: a mechanical device that insures proper synchronization of cross-patterning movements. The apparatus is manufactured by Flick-Reedy Education Enterprises of Bensonville, Illinois. It is a device 48 inches long and 14 inches wide. It has hand and knee pads that ride on rollers and are moved by muscular effort. These pads are interconnected by a system of cables and pulleys that literally force cross-patterning.

Treatment

Ss were pretested in July 1968, when they were enrolled in a pre-kindergarten clinic. Ss in the E group then trained on the Exer-Cor 3 minutes per day for a period of 3 months with their regular classroom teacher. Ss in the C group were not given any compensatory attention to control for the Hawthorne effect. In addition to the cross-patterning, the Ss were instructed to turn their heads in the direction of, and focus their eyes upon, the hand which was in the forward position. In addition to verbal instructions, Ss were given a demonstration of what was expected of them in the training exercises. At the end of the 3-month training period, they were posttested.

Tested Hypotheses

The specific statistical (null) hypotheses tested

are: (1) There is no difference between E and C groups in perceptuomotor skills as measured by the Purdue Perceptual Motor Survey. (2) There is no difference between E and C groups on the Draw A Man and Ten Dots subtests of the Anton Brenner Developmental Gestalt Test of School Readiness. (3) There is no difference between E and C groups in visual fusion as measured by the Keystone Visual Survey. The .05 level of significance was used for all hypotheses.

RESULTS

Hypothesis 1

A t-test for independent groups was performed, using difference (posttest minus pretest) scores for all subtests (hypothesis stated above).

Walking Board Subtest. The results showed that the control group was performing at a higher level than the experimental group. Results were not significant ($t(48) = -1.045$, $p > .05$).

Initiation of Movement Subtest. The experimental group scored higher than the control; results were not significant ($t(48) = .795$, $p > .05$).

Occular Pursuit. The control group scored significantly higher than the experimental group ($t(48) = -2.097$, $p > .05$).

Identification of Body Parts. The control group scored higher than the experimental group; results were not significant ($t(48) = -1.49$, $p > .05$).

Hypothesis 2

A t-test for independent data was performed using difference scores in both cases (the hypothesis is stated in the above section).

Draw A Man. The control group scored higher than the experimental group; results were not significant ($t(48) = -.510$, $p > .05$).

Ten Dot. The experimental group scored higher than the control group, but results were not significant ($t(48) = .790$, $p > .05$).

Hypothesis 3

There is no difference between E and C groups in visual fusion as measured by the Keystone Visual Survey. A chi-square test was performed on the differing frequencies, pre and post, of Ss exhibiting fusion problems. The results were not significant ($x^2 (48) = .07$, $p > .05$).

CONCLUSIONS AND DISCUSSION

That neurological training exercises, specifically cross-patterning exercises, will improve perceptuomotor skills as herein defined has not been shown by this study. Of the six measures of perceptuomotor skills, there were only three instances where the experimental group improved more than the control, and one case where the control group scored significantly higher than the experimental. This would seem to provide evidence against Delacato's theory of neurological training. However, in the present study there are several other possible explanations for the results.

One major weakness of the study was the use of intact classes. A much better design would have been to randomly select Ss for the E and C groups from within the same class; this would control for teacher differences. Unknown to E, the teacher assigned to the control group was conscious of the need for early childhood psychomotor training and handled her class accordingly. It is possible that she had a treatment effect that was much stronger than E's, especially since the experimental group only received 3 minutes of training daily. It would perhaps have been beneficial to start this at 3 minutes daily and gradually increase it. Or, perhaps even better, to have several treatment groups, each on a different time schedule.

Another possibility is that neurological training does not increase perceptuomotor skills. This is a source of controversy with the learned men arguing on both sides of the fence and each side backing their own position with their own research. The present study due to its inherent weaknesses does not offer evidence for either position. Carefully controlled research is necessary.

Future studies should be of truly random design and preferably a Solomon 4-group Design. This would eliminate the intact classroom problems and also enhance external validity. The treatment should, if at all possible, be administered by a trained person and care should be taken to prevent anything that can be adjudged to be a form of the same treatment or a like treatment.

REFERENCES

1. Alcuin, Sister N., "The Effect of Neurological Training on Disabled Readers," in Delacato, C., Neurological Organization and Reading, Thomas, Springfield, Illinois, 1966.

2. Barsch, R., "Project M.O.V.E. as a Model for Rehabilitation Theory," paper read at American Psychological Association Convention, Philadelphia, Pennsylvania, August 1963.

3. Betts, E.A., Foundations of Reading Instruction, American Book Company, New York, 1956.

4. Deboer, J.; Dallman, M., The Teaching of Reading, Holt, Rinehart, and Winston, New York, 1960.

5. DeHirsch, K.; Janskey, J.; Langford, W., Predicting Reading Failure, Harper and Row, New York, New York, 1966.

6. Delacato, C., The Treatment and Prevention of Reading Problems, Thomas, Springfield, Illinois, 1959.

7. Delacato, C., The Diagnosis and Treatment of Reading Problems, Thomas, Springfield, Illinois, 1963.

8. Delacato, C., Neurological Organization and Reading, Thomas, Springfield, Illinois, 1966.

9. Durrell, D., Improving Reading Instruction, World Book Company, New York, 1956.

10. Harris, A., How to Increase Reading Ability, Longmans, New York, 1956.

11. Kabot, R., "A Study of Improvement in Reading

Through Improvement of Neurological Organization," in Delacato, C., Neurological Organization and Reading, Thomas, Springfield, Illinois, 1966.

12. Mariam, Sister, O. P., "A Comparative Study of the Reading Disability in Neurologically Disorganized Fifth Grade Children," in Delacato, C., Neurological Organization and Reading, Thomas, Springfield, Illinois, 1966.

13. Masterman, J., "The Effect of Neurological Training on Reading Retardation," in Delacato, C., Neurological Organization and Reading, Thomas, Springfield, Illinois, 1966.

14. McClurg, W., "The Neurophysiological Basis of Reading Disabilities," The Reading Teacher, 22:615-621, 1969.

15. McGrath, Father F., "The St. Phillip Summer Remedial Course," in Delacato, C., Neurological Organization and Reading, Thomas, Springfield, Illinois, 1966.

16. McLeod, P., The Undeveloped Learner, Thomas, Springfield, Illinois, 1968.

17. Olson, N.; Olson, A.; Duncan, P., "Neurological Dysfunction and Reading Disability," The Reading Teacher, 22:157-162, 1968.

18. Painter, G., "The Effect of a Rhythmic and Sensory Motor Activity Program on Perceptual Motor Spatial Abilities of Kindergarten Children," Exceptional Children, 33:113, 116, 1966.

19. Robinson, H., Why Pupils Fail in Reading, University of Chicago Press, Chicago, Illinois, 1946.

20. Silver, A.; Hagin, R.; Hersh, M., "Reading Disability: Teaching Through Stimulation of Deficit Perceptual Areas," American Journal of Orthopsychiatry, 37:744-752, 1967.

21. Smith, D.; Carrigan, P., The Nature of Reading Disabilities, Harcourt, Brace, and Company, Chicago, Illinois, 1959.

EFFECT OF PRECISE OBJECTIVES UPON STUDENT ACHIEVEMENT IN HEALTH EDUCATION

GUS T. DALIS
Office of the Los Angeles County Superintendent of Schools

ABSTRACT

The purpose of this study was to determine what effect the communication of precise instructional objectives to students has on their learning. The study was designed (1) to provide data on whether student achievement can be influenced significantly by providing students, in advance of instruction, information on what is expected of them as an outcome of instruction and (2) to investigate various ways of communicating to students, in writing, that which is to be learned in class. The Ss for this study were selected from five tenth-grade health and safety classes taught by the same teacher. Of the 143 Ss, one third in each class was randomly assigned to one of three treatment groups. For treatment groups one through three, the participants received precisely stated instructional objectives, vaguely stated instructional objectives, and short paragraphs of health information, respectively. Ss receiving prior to instruction precise information on what is expected of them showed greater achievement than those who received vague or related information.

DESPITE emerging pronouncements as to the value and utility of instructional objectives to the teaching-learning situation, many teachers and curriculum workers still look upon objectives as necessary decorations to satisfy the curriculum theorist but beyond that objectives serve no useful purpose. With the existence of such a situation, perhaps there is a need to consider not only how instructional objectives should be stated but also the way they might best be utilized in the teaching-learning setting so as to favorably influence student achievement. Ultimately, the value of objectives to teachers will be the degree to which these statements serve a useful purpose in the teaching-learning process.

A study was designed to determine the effect communication of precise instructional objectives to students has on their learning. The study was conducted (1) to provide data on whether student achievement can be influenced significantly by providing students, in advance of instruction, information on what is expected of them as an outcome of instruction and (2) to investigate various ways of communicating to students, in writing, that which is to be learned in class.

RELATED LITERATURE

Although there have been no previous studies showing the effectiveness of providing learners with precise instructional objectives prior to instruction in public settings: (1) the need for such studies, while not expressed by authorities in the literature, seems to exist in that some in the education profession allude to the usefulness of learners having knowledge of classroom instructional objectives (8, 12, 13:36, 15:91, 17, 20, 23); (2) studies in other settings have indicated a usefulness in providing learners with instructional objectives in advance of instruction (10, 13:28, 20). For instance, it has been found that knowledge of objectives by adults reduces the time for educating them in tasks related to their jobs, and objectives possibly increase achievement by college students; (3) studies that assess the effect of using behavioral instructional objectives with teachers revealed enhanced achievement by the learners taught by these teachers (21, 24); and (4) other related studies where instructions and advance organizers were used with learners prior to instruction it was found that their achievement was influenced favorably (1:235, 2, 3, 4, 5, 9, 11:268, 19:640-641, 29:175).

Authorities have suggested that giving learners objectives will help them know what is expected of them in advance of instruction on the basis that such a teaching strategy: (1) helps the learner identify the required terminal performance (12:357-358); (2) assists the learner in maintaining his own control of learning task reinforcement (13:26-27); (3) provides knowledge of goals to attain, which in turn is both instructive and motivating (14); (4) facilitates

the exploration of alternatives for learning and provides direction for this exploration (6:43-44); (5) results in a greater commitment by learners to ends and the means to those ends (16:522-523, 18:298-299, 25:332-334, 26); (6) helps learners discriminate between relevant and irrelevant learning material (1:85-86, 2:270-271); (7) predisposes learners toward certain kinds of behavior rather than other kinds (11:269); (8) actively engages learners in utilizing their prior knowledge (28); (9) facilitates the learner's organization of relevant knowledge which will direct his thinking to the learning task (11:6 and 275); (10) provides a motive for specific rather than random learner behavior (22:16 and 62); and (11) motivates learners and that indirectly this motivation results in increased learner effort, attention, and readiness to learn (1:228 and 235). Whereas there is much speculation concerning the value of giving learners objectives prior to instruction there are few studies supporting such an approach and none conducted in the public schools.

METHOD
Subjects

A teacher with five health and safety classes in a predominately middle and upper-middle class high school was selected as the study teachers. Within the five classes there were a total of 143 tenth-grade students. At the beginning of the study each of the S's was randomly assigned to one of three treatment groups within each of the five classes. The assignment to a treatment group was conducted by using a table of random numbers.

Materials

As part of the study, the teacher was asked to conduct a 3 week unit on growth and development within the health education program. None of the content in the unit was presented in the high school course of study prior to the time of this study. The unit was developed in accordance with the School Health Education Study (SHES) (27:42-45) concept "growing and developing follows a predictable sequence, yet is unique for each individual." The first four of the five unit objectives identified by the SHES at the high school level for this concept served as the framework for the development of the teaching unit. The unit was designed as a comprehensive teaching "package" for teacher use and included an identification of content related to each unit objective and a variety of learning opportunities keyed to each unit objective.

From each unit objective a number of related precise and vague instructional objectives for student use were developed. The precise instructional objectives contained explicit specific content, the kind of overt behavior expected of the learner with respect to this content, conditions to be imposed upon the learner when he is demonstrating mastery of the objectives, and the inclusion of what will be acceptable performance. The vague instructional objectives were somewhat similar to the precise instructional objectives except that both the objective content and behavior dimensions were general. In the vague objectives the content was presented in broad and general language. The behavioral terms used in these objectives were open to more interpretations than the terms used with the precise objectives. Also, the vague objectives did not contain a statement of conditions to be imposed upon the learner when he is

demonstrating his attainment of the objective, nor was there an indication of what would be acceptable learner performance.

A total of sixteen precise and vague instructional objectives was developed from the four teaching unit objectives. Both the precise and vague instructional objectives represent a sample of a population of such objectives that could have been developed and that were implicit within the four unit objectives. In addition, sixteen separate short paragraphs of written health information unrelated to classroom learnings were developed to serve as a placebo with the control group of students. Each of the paragraphs of written health information, precise instructional objectives, and vague instructional objectives was placed on a separate sheet of paper. These sheets of paper were referred to as "messages" whenever they were discussed with Ss.

For each precise instructional objective one multiple choice test item was developed to assess the student's understanding of the objective. The same test items were used with the respective vague instructional objectives since each item was but a sample of many possible test items that could have been prepared for these more general objectives. Therefore, each "message" containing either a precise or vague instructional objective included a relevant multiple choice test item. The instructions on each "message" sheet directed the S to select from an array of four choices the one that best went with their objective. For the written information used as the placebo a time-consuming activity comparable to the test items for the precise and vague instructional objectives was developed to accompany this information.

A sixty-eight item criterion test was developed to assess student achievement at the conclusion of the growth and development unit. Also, an opinionnaire was designed to secure certain reactions by participants during the study.

Procedures

In conducting this study the Posttest-Only Control Group Design (7:195-197) was employed. A comparison of student achievement was conducted among those Ss provided with precise instructional objectives (group one), those provided with a set of vague instructional objectives (group two), and those provided with a placebo (group three) in advance of instruction.

At the beginning of the study Ss were told that they had been selected to participate in an experiment to determine whether written messages to be given them periodically during a 3-week period would be of any assistance in their classwork. They were informed that different people in the class would be receiving different messages and that in order for the experiment to work it was necessary to maintain absolute secrecy. During this introductory discussion the topic of "experiments" was discussed including the importance of "maintaining certain controls in order for experiments to work." This discussion along with the announcement that the Ss would participate in an experiment whereby different people would have different information was pursued with the intent that such an approach of honesty would enhance student compliance with the experiment procedures. Initially,

Ss were told that their grade would not be affected by their different messages and that they could withdraw from the experiment at any time without penalty to their class standing. None of the students elected to withdraw from the study. During the 3-week unit, the teacher would pause at points indicated in the teaching unit plan and provide each student with his appropriate message. The messages were prepared in advance by the investigator with the student's name on each folded message sheet. Throughout the experiment the teacher remained unaware of the specific character of the information being given the Ss.

On the last day allocated for the experiment the criterion test was administered to evaluate achievement of material contained in the growth and development unit. Also, each S responded to an "anonymous" opinionnaire coded in such a fashion whereby it was possible through a classroom seating chart to identify each respondent. One opinionnaire question concerned the amount of study time spent outside of class each day during the study of growth and development. A second concerned the gaining of any information about messages given to other students. The test data from six Ss who indicated they had gained information about messages from someone else were not included in the analysis. These six Ss were evenly distributed among the three treatment groups. In addition four other Ss were not included in the study. Three dropped out of school prior to the conclusion of the study, and one was a native speaker of Spanish who spoke and read very little English.

RESULTS

Contained in Table 1 are data on the analysis of variance of the criterion test dependent variable between treatment groups. The very high F ratio, 10.809, clearly shows that there was a treatment effect favoring group one, the group presented with precise instructional objectives prior to instruction. Therefore, the hypothesis indicating that the group presented with precise instructional objectives prior to instruction will demonstrate greater achievement than will the group presented with vague instructional objectives is accepted beyond the .99 level of confidence.

The means on the criterion test scores of treatment groups one through three were 40.4, 31.4, and 32.1 respectively. The mean of 40.4 by the group presented with precise instructional objectives is significantly separated from the mean of 31.4 by the group presented with vague instructional objectives and the mean of 32.1 by the group presented with the placebo. Thus, the hypothesis that indicated that the group presented with either

precise or vague instructional objectives prior to instruction will demonstrate greater achievement than will the group presented with a placebo can be assumed to be rejected. Major significant differences are due to the high level of achievement of group one on the criterion test.

Group one obtained a mean of 8.9 on the instructional objective understanding test. The standard deviation for this group was 4.6. The vague instructional objective group (group two) obtained a mean of 2.2 with a standard deviation of 1.6 on the instructional objective test. The mean difference between groups one and two was 6.7. When computed, the value for t was equal to 67.25, significant at the .999 level of confidence. Since there was an extremely high t value, which shows the significance of differences between groups one and two, there is probably a real difference between these groups on instructional objective understanding. Due to the marked differences in standard deviation between the groups caution is warranted on accepting the extremely high t value. The hypothesis can thus be accepted that the precise instructional objective group will be more able to select activities that go with their objectives than will the vague instructional objective group.

By using the Kuder-Richardson formulas 20 and 21, it was found that the reliability for groups one and two was .90 on the instructional objective understanding test. This very high value indicates that the test as a whole, regardless of the treatment group, is particularly reliable, and internally consistent. A .90 reliability coefficient is quite large for a 16-item test.

By the way of description, less overall average time was spent in studying outside of class by group one than by groups two and three. Group one actually spent an average of 17.4 minutes, group two 27.0 minutes, and group three 20.4 minutes studying daily outside of class. This average amount of time spent studying among the three groups, however, was not significantly different at the .99 level of confidence.

CONCLUSIONS

According to the findings in this study it was possible to enhance health education classroom achievement by using precise instructional objectives in advance of instruction with high-school-age learners. These objectives, however, must be precisely stated otherwise their value to learning efficiency is doubtful. In fact, instructional objectives that are vaguely stated and are general both in content and behavior may deter learner achievement when given to him prior to instruction.

Evidence from this study supports the idea that individuals with precise instructional objectives were quite able to select activities related to these objectives. Whereas, those individuals guided by vague instructional objectives seemingly became confused and were unable to select activities that related to their objectives. Apparently the vague objectives did not provide the necessary direction and information needed to facilitate the matching of relevant activities to instructional objectives.

The study findings revealed that the precision of stating instructional objectives did not affect, in one

TABLE 1

ANALYSIS OF VARIANCE OF CRITERION TEST AMONG DEPENDENT VARIABLE TREATMENT GROUPS

Source	Sum of Squares	df	MS	F
Between	2106.170	2	1053.085	10.978*
Within	12470.151	130	95.924	
Total	14576.321	132		

*p .01
F .01 (2, 130)=4.79

way or another, the amount of time spent studying daily outside of class by those learners being guided by these objectives.

REFERENCES

1. Ausubel, David P., The Psychology of Meaningful Verbal Learning, Grune and Stratton, Inc., New York, 1963.

2. Ausubel, David P., "The Use of Advance Organizers in the Learning and Retention of Meaningful Verbal Material," Journal of Educational Psychology, 51:267-272, October 1960.

3. Ausubel, David P.; Fitzgerald, Donald, "Organizer, General Background, and Antecedent Learning Variables in Sequential Verbal Learning," Journal of Educational Psychology, 53: 243-249, December 1962.

4. Ausubel, David P.; Fitzgerald, Donald, "The Role of Discriminability in Meaningful Verbal Learning and Retention," Journal of Educational Psychology, 52:266-274, October 1961.

5. Ausubel, David P.; Youssef, Mohamed, "Role Discriminability in Meaningful Parallel Learning," Journal of Educational Psychology, 54: 331-336, December 1963.

6. Bruner, Jerome S., Toward a Theory of Instruction, The Belknap Press of Harvard University Press, Cambridge, Massachusetts, 1966.

7. Campbell, Donald T.; Stanley, Julian C., "Experimental and Quasi-experimental Designs for Research on Teaching," Handbook of Research on Teaching, Rand McNally and Company, Chicago, 1963.

8. Canfield, Albert A., "A Rationale for Performance Objectives," Audiovisual Instruction, 13:127-129, February 1968.

9. Dawson, Kenneth E., "The Effectiveness of Subsuming Concepts in Teaching Industrial Arts," unpublished doctoral dissertation, University of Maryland, 1965.

10. Dressel, Paul L., Evaluation in General Education, William C. Brown Company, Dubuque, Iowa, 1954.

11. Forgus, Ronald H., Perception: The Basic Process in Cognitive Development, McGraw-Hill Book Company, New York, 1966.

12. Gagné, Robert M., "The Acquisition of Knowledge," Psychological Review, 69:355-365, July 1962.

13. Gagné, Robert M., "The Analysis of Instructional Objectives for the Design of Instruction," in Glaser, Robert, (ed.), Teaching Machines and Programed Learning, II, Data and Directions, National Education Association, Washington, D. C., 1965.

14. Glaser, Robert, "Objectives and Evaluation: An Individualized System," Science Education News, American Association for the Advancement of Science, 1-3, June 1967.

15. Haberman, Martin, "Behavioral Objectives: Bandwagon or Breakthrough," The Journal of Teacher Education, 19:91-94, Spring 1968.

16. Jackson, Philip W.; Strattner, Nina, "Meaningful Learning and Retention: Noncognitive Variables," Review of Educational Research, 34: 513-529, December 1964.

17. Kapfer, Phillip, "What Should Be Done to Improve the Curriculum," PACE Report, October 1967.

18. Lewin, Kurt; Lippitt, Ronald; White, Ralph K., "Patterns of Aggressive Behavior in Experimentally Created Social Climates," The Journal of Social Psychology, 10:271-299, May 1939.

19. Lumsdaine, A. A., "Instruments and Media of Instruction," in Gage, N. L., (ed.), The Handbook of Research on Teaching, Rand McNally and Company, Chicago, 1963, Chapter 12.

20. Mager, Robert F.; McCann, John, Learner-Controlled Instruction, Varian Associates, Palo Alto, California, no date.

21. McNeil, John D., "Concomitants of Using Behavioral Objectives in the Assessment of Teacher Effectiveness," The Journal of Experimental Education, 36:69-74, Fall 1967.

22. Miller, George A.; Galanter, Eugene; Pribram, Karl H., Plans and the Structure of Behavior, Henry Holt and Company, New York 1960.

23. Miller, R. B., "Task Description and Analysis," Gagné, R. M., (ed.), in Psychological Principles in System Development, Holt, Rinehart and Winston, New York 1962, pp. 187-228.

24. Moffett, George M., "Use of Instructional Objectives in the Supervision of Student Teachers," unpublished doctoral dissertation, University of California, Los Angeles, 1966.

25. Raven, Bertram, "The Dynamics of Groups," Review of Educational Research, 29:332-343, October 1959.

26. Raven, Bertram H.; Rietsma, Jan, "The Effects of Varied Clarity of Group Goal and Group Path Upon the Individual and His Relation to His Group," Human Relations, 10:29-47, February 1957.

27. School Health Education Study, Health Education: A Conceptual Approach to Curriculum Design, School Health Education Study, Washington, D. C., 1967.

28. Whittrock, M. C., "Effects of Certain Sets Upon Complex Verbal Learning," Journal of Educational Psychology, 54:85-88, April 1963.

29. Whittrock, M. C., "Set Applied to Student Teaching," Journal of Educational Psychology, 53:175-180, August 1962.

SOME EVIDENCE CONCERNING THE VALIDITY

OF AN ELEMENTARY SCHOOL FORM

OF THE DOGMATISM SCALE

DONALD W. FELKER
DONALD J. TREFFINGER
Purdue University

ABSTRACT

Five hypotheses concerning the validity of Figert's elementary school form of Rokeach's dogmatism scale were investigated. Pupils (N=120) from fourth, fifth, and sixth grade classes participated in the study. The results offered little support for the validity of the Figert test. Only one of the five hypotheses was supported, and that only partially. It seems necessary, therefore, to conclude that scores on this test should not be interpreted as an assessment of dogmatism unless other evidence can be obtained to provide support for the test's validity, or an acceptable alternative interpretation of these data can be formulated.

THE PURPOSE of this study was to investigate the validity of an elementary school form of a dogmatism scale which Figert (4) presented as an adaptation of Rokeach's (11) Dogmatism Scale. Figert concluded that his scale was

functioning relatively effectively as a measuring device and was measuring some of the same facets of openmindedness-closedmindedness among children that adult forms of the instrument measure among adults (4:20-21).

Some of these data, he argued, could be interpreted as evidence for the validity of the instrument. Such data were:

(1) Mean scores for pupils in grade 4 were significantly greater than means for pupils in grades 5 and 6;

(2) There was a tendency toward an inverse relationship between test scores and socioeconomic class (SES) indices, although means did <u>not</u> differ significantly among SES levels;

(3) Means for parochial school pupils were significantly greater than means for pupils in two (of four) public schools studied.

Figert (4:21) also suggested that the scale should be used for some scale-validation studies following the techniques developed by Rokeach and others.

The present study is such a scale-validation study. In addition to reexamining, in a different sample, differences among grade levels in scores on Figert's test, the present study tested several other hypotheses concerning dogmatism among elementary school pupils. These hypotheses were derived from relationships previously established with older Ss.

Rokeach (11) presented evidence, for example, to support the prediction of a negative relationship between self-concept and dogmatism. Holden (7) reported that external locus of control or responsibility was positively correlated with scores on the California F-Scale. Many descriptions of the correlates of creativity (1, 5) have suggested that highly creative persons are tolerant of ambiguity, open to experience, confident, self-assertive, and independent in judgment. Such a description is similar in many ways to Rokeach's (11) description of the openminded individual. Rokeach (12) argued that openmindedness may be a prerequisite for creativity; evidence suggesting a negative relationship between dogmatism and creativity was presented by Jacoby (8). Mouw (9) also presented evidence to support the prediction of a negative relationship between dogmatism and complex cognitive abilities.

Thus, the following hypotheses were formulated, which, on the basis of the previous work cited, should provide evidence concerning the construct validity of Figert's test:

If the Elementary School Form of the Dogmatism Scale (4) measures dogmatism, then:

(1) There will be differences in mean scores among several grade levels;

(2) There will be a negative relationship between dogmatism test scores and scores on the Piers-Harris self-concept scale (10);

(3) There will be a negative relationship between dogmatism test scores and assessment of internal locus of control, as measured by the Intellectual Achievement Responsibility Questionnaire (3);

(4) There will be a negative relationship between scores on the dogmatism test and scores on measures of creative problem solving abilities, assessed using a battery developed by Treffinger and Ripple (13, 14);

(5) There will be a negative relationship between dogmatism test scores and attitudes about creative thinking, and between scores on the dogmatism test and self-concept of creative problem solving ability, as measured by the Childhood Attitude Inventory for Problem Solving (2).

METHOD

Sample

Fourth, fifth, and sixth grade classes from public schools in northern Indiana (N=120) participated in this study.

Instruments

In addition to Figert's Elementary School Form of the Dogmatism Scale, all pupils were given several other instruments. These were:

(a) The Intellectual Achievement Responsibility (IAR) questionnaire, which was developed by Crandall, Katkovsky, and Crandall (3); (b) The Piers-Harris Self-Concept Scale (P-H), developed by Piers and Harris (10); (c) A General Problem Solving battery (GPS), developed by Treffinger and Ripple (13, 14); and (d) The Childhood Attitude Inventory for Problem Solving (CAIPS), developed by Covington (2).

The reliability and validity of these instruments have been discussed in the sources indicated; it was concluded that, for the purposes of this study, sufficient evidence was available to warrant their utilization. All tests were administered by classroom teachers, using standardized directions, and scored by trained graduate students.

Analyses

Intercorrelations among all variables were computed for the entire sample, as well as separate matrices for boys and girls. Then, grade level differences in scores on Figert's dogmatism test were examined, using one-way analysis of variance (6) to compare means among fourth, fifth, and sixth grades. Next, extreme groups on the dogmatism scale (40 highest scoring pupils and 40 lowest) were formed. These groups were compared on P-H,

CAIPS, and IAR scores. Finally, for a sample of forty-three pupils on whom identifiable problem solving data were available, correlations were computed between GPS scores and dogmatism scores.[1] The alpha level was set at .05 for all ANOVA's and tests of the significance of correlation coefficients.

RESULTS

Hypothesis one was tested using one-way analysis of variance of dogmatism scores among grade levels. The means for grade 4 (104.37), grade 5 (105.00), and grade 6 (106.30) did not differ significantly ($F \leq 1$, with 2,119 df). It was necessary, therefore, to conclude that hypothesis one was not supported.

Correlation coefficients, to test hypotheses two and five, are presented in Table 1. Among the eighteen coefficients, only one was significantly different from zero: the correlation between dogmatism scores and pupils' attitudes about creative thinking and problem solving (CAIPS, I) was -.271. This correlation was in the direction predicted by our hypothesis: pupils with higher dogmatism scores tended also to have lower scores on the attitude measure.

In order to examine the predicted relationships more closely, differences between pupils in the highest third of the distribution of dogmatism scores and pupils in the lowest third were examined. These results are summarized in Table 2. Because of the limited sample available, GPS scores were not included in these comparisons.

TABLE 1

CORRELATIONS BETWEEN DOGMATISM SCORES AND SELF-CONCEPT, LOCUS OF CONTROL, CREATIVE PROBLEM SOLVING, AND ATTITUDES FOR BOYS, GIRLS, AND TOTAL SAMPLE

Variable	Boys	N	Girls	N	Total	N
P-H	-.058	60	-.024	60	-.037	120
IAR+	-.094	60	.010	60	-.048	120
IAR-	-.135	60	-.202	60	-.168	120
GPS	.048	23	.235	20	.125	43
CAIPS I	-.271*	60	-.101	60	-.221	120
CAIPS II	-.228	60	-.114	60	-.127	120

*$p < .05$

The only significant difference between high and low dogmatism groups was on part I of CAIPS (attitudes about creative thinking and problem solving). Pupils in the low dogmatism group had significantly higher scores on the attitude measure than pupils in the high dogmatism group.

DISCUSSION AND CONCLUSIONS

The purpose of this study was to test several hypotheses which, if supported, would provide evidence for the validity of the elementary school form of the dogmatism scale, developed by Figert (4). Five hypotheses were tested; the results for each will be discussed.

TABLE 2

DIFFERENCES BETWEEN THE UPPER AND
LOWER THIRDS (N = 40) WHEN RANKED
ON FIGERT'S TEST SCORES

Variable	MS_B	MS_W	F^1	p
P - H	.01	166.69	< 1	n. s.
IAR +	.45	4.25	< 1	n. s.
IAR -	6.61	7.73	< 1	n. s.
CAIPS I	68.45	13.15	5.206	$p < .05$
CAIPS II	19.01	23.13	< 1	n. s.

^{1} With 1 and 78 df for all variables.

Hypothesis one, which dealt with differences among grade levels on the dogmatism test, was not supported. Our results, therefore, do not substantiate those reported by Figert (4).

Hypothesis two, which proposed a negative relationship between dogmatism scores and scores on the Piers-Harris self-concept scale, was not supported. The correlation between these tests in our sample was not reliably different from zero; nor were there significant self-concept differences between groups high and low on dogmatism.

Hypothesis three predicted a negative relationship between dogmatism test scores and a measure of internal locus of control. Again, there was no support for this hypothesis, since the correlations obtained did not differ significantly from zero, nor did high and low dogmatism groups differ significantly on IAR scores.

Hypothesis four predicted a negative correlation between dogmatism scores and creative problem solving scores. Since the correlation obtained did not differ significantly from zero, no support was found for the hypothesis.

For the predicted negative relationship between dogmatism scores and attitudes and self-concept about creative thinking and problem-solving (hypothesis five), limited support was found. Pupils who had lower dogmatism scores tended to have higher (more favorable) attitudes toward creative thinking and problem solving. There were no significant relationships, however, between dogmatism scores and pupils' expression of self-concept of creative problem-solving ability.

It would appear that our results offer virtually no support for the validity of the Figert test. Among five predictions, we have found significant support for only one, and even in that case, the support was limited.

We must conclude that these data cast serious doubt on the usefulness of the Figert test. Unless other evidence can be obtained to provide support for the test's validity or to provide an acceptable alternative interpretation of our data, it seems necessary to conclude that scores on this test should <u>not</u> be interpreted as an assessment of dogmatism, as this construct has been defined elsewhere in psychological research.

FOOTNOTE

1. Because of an error in directions some pupils did not put their names on the problem solving tests.

REFERENCES

1. Barron, Frank, <u>Creative Person and Creative Process</u>, Holt, Rinehart and Winston, New York, 1969.

2. Covington, M. V., "A Childroon Attitude Inventory for Problem Solving," Berkeley Creativity Project, Berkeley, California, 1967.

3. Crandall, V. C.; Katkovsky, W.; Crandall, V. J., "Childrens' Beliefs in Their Own Control of Reinforcement in Intellectual-Academic Achievement Situations," <u>Child Development</u>, 36:91-109, 1965.

4. Figert, R. S., Jr., "An Elementary School Form of the Dogmatism Scale," <u>Journal of Experimental Education</u>, 37:19-23, 1968.

5. Golann, S. E., "Psychological Study of Creativity," <u>Psychological Bulletin</u>, 60:(no. 6)548-565, 1963.

6. Guenther, W. C., <u>Analysis of Variance</u>, Prentice-Hall, Englewood Cliffs, 1964.

7. Holden, K., <u>Attitude Toward External Versus Internal Control and Learning of Reinforcement Sequences</u>, unpublished master's thesis, Ohio State University, Columbus, 1958.

8. Jacoby, J., "Open Mindedness and Creativity," <u>Psychological Reports</u>, 20:822, 1967.

9. Mouw, J. T., "Effect of Dogmatism on Levels of Cognitive Processes," <u>Journal of Educational Psychology</u>, 60:(no. 5)365-369, 1969.

10. Piers, E. V.; Harris, D. B., "Age and Other Correlates of Self-Concept in Children," <u>Journal of Educational Psychology</u>, 55:91-95, 1964.

11. Rokeach, M., <u>The Open and Closed Mind</u>, Basic Books, New York, 1960.

12. Rokeach, M., "In Pursuit of the Creative Process," Steiner, G. A. (ed.), <u>The Creative Organization</u>, University of Chicago, Chicago, 1965, pp. 66-88.

13. Treffinger, D. J.; Ripple, R. E., "The Effects of Programmed Instruction in Productive Thinking on Verbal Creativity and Problem Solving Among Elementary School Pupils," final report of USOE Research Project OEG-0-8-080002-0220-010, Cornell University, Ithaca, New York, 1968.

14. Treffinger, D. J.; Ripple, R. E., "The Effects of Programed Instruction in Productive Thinking on Creative Problem Solving Abilities and Attitudes Among Pupils in Grades Four Through Seven," <u>Irish Journal of Education</u>, 1970 (in press).

TEACHER-PRINCIPAL RELATIONSHIPS

IN "HUMANISTIC" AND "CUSTODIAL"

ELEMENTARY SCHOOLS[1]

WAYNE K. HOY
Rutgers University

JAMES B. APPLEBERRY
Oklahoma State University

ABSTRACT

Following the lead of earlier research on pupil control, the concepts of "humanism" and "custodialism" were used to refer to contrasting types of individual ideology and the types of school organization they seek to rationalize and justify. The Pupil Control Ideology Form (PCI) and the Organizational Climate Description Questionnaire (OCDQ) were personally administered by a researcher to virtually all the professional personnel of forty-five elementary schools; from this sample, fifteen "humanistic" and fifteen "custodial" schools were identified. Using analysis of variance procedures, comparisons between the patterns of social interactions of professional staff in humanistic and custodial schools revealed statistically significant differences on four of the eight dimensions of the OCDQ. In addition, as predicted, humanistic schools were significantly more "open" than custodial schools. The results suggested that the pupil control orientation of a school may provide another important step in identifying the "social climate" of the school.

CONTROL IS a problem faced by all organizations, but it is especially important in service organizations which work with people or clients rather than material goods. Public schools are social units specifically vested with a service function, the socialization of the young. Furthermore, they are a type of service organization in which neither the organization nor the client exercises choice concerning participation in the relationship; that is, public schools have no choice in the selection of clients (students), and the client must (in the legal sense) participate in the organization (3). It should not be surprising that organizations of this type are likely to be confronted with some clients who have little or no desire for the services of the organization, a factor which accentuates the problem of client control.

Indeed, there is no lack of opinion or prescription on pupil control in public schools, but unfortunately there is little systematic study on the subject, much less, study which begins from the perspective of the school as a social system. Studies which have focused on the school as a social system have described antagonistic student subcultures and attendant control problems (4, 6, 12). For example, Waller's (12) classic analysis of the social organization of the school underscored the importance and centrality of pupil control in both the structural and normative aspects of the school culture.

More recent studies of public schools also have emphasized the saliency of pupil control in the organizational life of schools (9, 10, 11, 13, 14). For example, in one study pupil control was described as the "integrative theme" of the school which gave meaning to teacher-teacher and teacher-administrator relations. In the words of the researchers, "While many other matters influenced the tone of the school, pupil control was a dominant motif" (14:107).

Schools differ in terms of the nature of their educational viewpoints and policies concerning control of students. Some schools are characterized by stress on maintenance of order, impersonality, distrust of students, and, in general, a punishment-centered orientation toward students. Other schools are marked by an accepting, trustful view of students, and confidence in students to be self-disciplining and responsible. Given the apparent significance of pupil control, to what extent are these kinds of differences in pupil control orientation related to other important characteristics of schools? This question led to the hypothesis that teacher-teacher and teacher-principal interactions would be significantly different in schools with humanistic pupil control orientation than in schools with a custodial orientation.

"CUSTODIALISM" AND "HUMANISM"

Following the lead of earlier research on pupil

control, the concepts of "humanism" and "custodi-alism" were adopted to refer to contrasting types of individual ideology and the types of school organization that they seek to rationalize (13).

In its extreme form, the custodial orientation favors a rigid and highly controlled setting concerned primarily with the maintenance of order. Students are stereotyped in terms of their appearance, behavior, and parents' social status. Teachers who hold a custodial orientation conceive of the school as an autocratic organization with a rigid pupil-teacher status hierarchy; the flow of power and communication is unilateral downward. Students must accept the decisions of teachers without question. Student misbehavior is viewed as a personal affront; students are perceived as irresponsible and undisciplined persons who must be controlled through punitive sanctions. Impersonality, pessimism, and "watchful mistrust" imbue the atmosphere of the custodial school.

The model for the humanistic orientation, on the other hand, is the school conceived as an educational community in which students learn through cooperative interaction and experience. Learning and behavior are viewed in psychological and sociological terms rather than moralistic ones. Self-discipline is substituted for strict teacher control. The humanistic orientation leads teachers to desire a democratic atmosphere with its attendant flexibility in status and rules, sensitivity to others, open communication, and increased student self-determination. Both teachers and pupils are willing to act on their own volition and to accept responsibility for their actions.

TEACHER-PRINCIPAL INTERACTIONS

In a major study of seventy-one elementary schools, Halpin and Croft (8) identified and described eight basic characteristics of social interaction between the principal and the teachers. Four of the characteristics refer to teacher behavior: Disengagement, Hindrance, Esprit, and Intimacy; and four describe principal behavior: Aloofness, Production Emphasis, Thrust, and Consideration. The behavior described by each characteristic is briefly described below (7):

Disengagement indicates that the teachers do not work well together. They pull in different directions with respect to the task; they gripe and bicker among themselves.

Hindrance refers to the teachers' feeling that the principal burdens them with routine duties, committee demands, and other requirements which they construe as unnecessary busy-work.

Esprit refers to "morale." The teachers feel that their social needs are being satisfied, and that they are, at the same time, enjoying a sense of accomplishment in their job.

Intimacy refers to the teachers' enjoyment of friendly social relations with each other.

Aloofness refers to behavior by the principal which is characterized as formal and impersonal. He "goes by the book" and prefers to be guided by rules and policies rather than to deal with the teachers in an informal, face-to-face situation.

Production Emphasis refers to behavior by the principal which is characterized by close supervision of the staff. He is highly directive and task-oriented.

Thrust refers to behavior marked not by close supervision of the teacher, but by the principal's attempt to motivate the teachers through the example which he personally sets. He does not ask the teachers to give of themselves anything more than he willingly gives of himself; his behavior, though starkly task-oriented, is nonetheless viewed favorably by the teachers.

Consideration refers to behavior by the principal which is characterized by an inclination to treat the teachers "humanly," to try to do a little something extra for them in human terms.

In addition, Halpin and Croft (8) conceptualized social interactions of professional personnel of schools in terms of a more general factor, openness. The openness of a school refers to actions which emerge freely and without constraint; that is, the behavior of the group members is genuine or authentic. Leadership acts are readily initiated from both the principal and teachers, and the group is not inordinately concerned with either task achievement or social-needs satisfaction. Satisfaction on both counts emerges easily and almost effortlessly.

The concept of openness in organizational behavior seems highly compatible with a humanistic pupil control orientation. If pupil control is a salient feature of the organizational life of schools, it seems reasonable to further hypothesize that "humanistic" schools will be significantly more open in teacher-principal interactions than "custodial" schools.

PROCEDURES

Instruments

The PCI Form was the operational measure for pupil control orientation; it consists of twenty Likert-type items. Responses are scored from 5 (strongly agree) to 1 (strongly disagree): the higher the overall score, the more custodial the ideology of the respondent.

Examples of items used include: "A few pupils are just young hoodlums and should be treated accordingly." "It is often necessary to remind pupils that their status in schools differs from that of teachers." And, "Pupils can be trusted to work together without supervision" (score reversed).

In earlier research (13), split-half reliability coefficients of the instrument in two samples were .95(N=170) and .91(N=55) with application of the Spearman-Brown Formula. Validity of the measure was supported by principals' judgments of the ideology of certain of their teachers. Teachers judged to be most custodial by their principals had significantly higher (p <.01 using t-test procedures) PCI Form scores than a like number of teachers judged to be most humanistic. [2]

The OCDQ is composed of sixty-four Likert-type items which teachers and principals may use to describe various aspects of social interaction in their schools. By factor analysis, Halpin and Croft (8) subdivided the OCDQ into eight dimensions, each with a corresponding subtest. The Disengagement, Hindrance, Esprit, and Intimacy subtests refer primarily to the behavior of the teachers; and the Aloofness, Production Emphasis, Thrust, and Consideration subtests to the behavior of principals. Further factor analysis of school profiles led to the identification of a general openness factor. Openness scores for schools can be computed by summing the Esprit and Thrust subtest scores and then subtracting the Disengagement score.

Findings of numerous studies have supported the validity and reliability of the eight OCDQ subtests (1, 2). For example, a major validity study was conducted by Andrews (1); using the method of construct validity, he concluded ". . . the subtests of the Organizational Climate Description Questionnaire provide reasonably valid measures of important aspects of the school principal's leadership in perspective of interaction with his staff" (1:333).

Sample

Forty-five elementary schools in thirty school districts comprised the sample of this study. Several criteria were used in the selection of elementary schools for study. In order to allow sufficient opportunity for the development of interaction patterns between the principal and teachers, only schools with principals who were near the completion of at least their second year as full-time principals and who served in only one building were included in the sample. Further, elementary schools were selected from various types of communities: rural, town or small city, suburban, and urban.

Originally, fifty schools seemed to meet the selection criteria and were asked to participate in the study. Four schools declined the invitation to participate, and further information excluded another school from the sample. The forty-five elementary schools that agreed to participate were personally visited by a researcher and both the PCI and OCDQ were administered to the professional personnel during regularly scheduled faculty meetings. Virtually all of the teachers and principals in each school responded to the instruments.

In this phase of the investigation, fifteen relatively "custodial" and fifteen relatively "humanistic" elementary schools were identified from the original group of forty-five. Those schools with the highest mean PCI scores were designated as custodial schools (range = 52.2 - 61.8) while those schools with the lowest PCI scores were termed humanistic schools (range = 45.7 - 52.5).

RESULTS

As predicted, the examination of the profiles of humanistic and custodial schools found in Figure 1

FIGURE 1

A COMPARISON OF PROFILES OF MEAN SUBTEST SCORES OF HUMANISTIC AND CUSTODIAL ELEMENTARY SCHOOLS
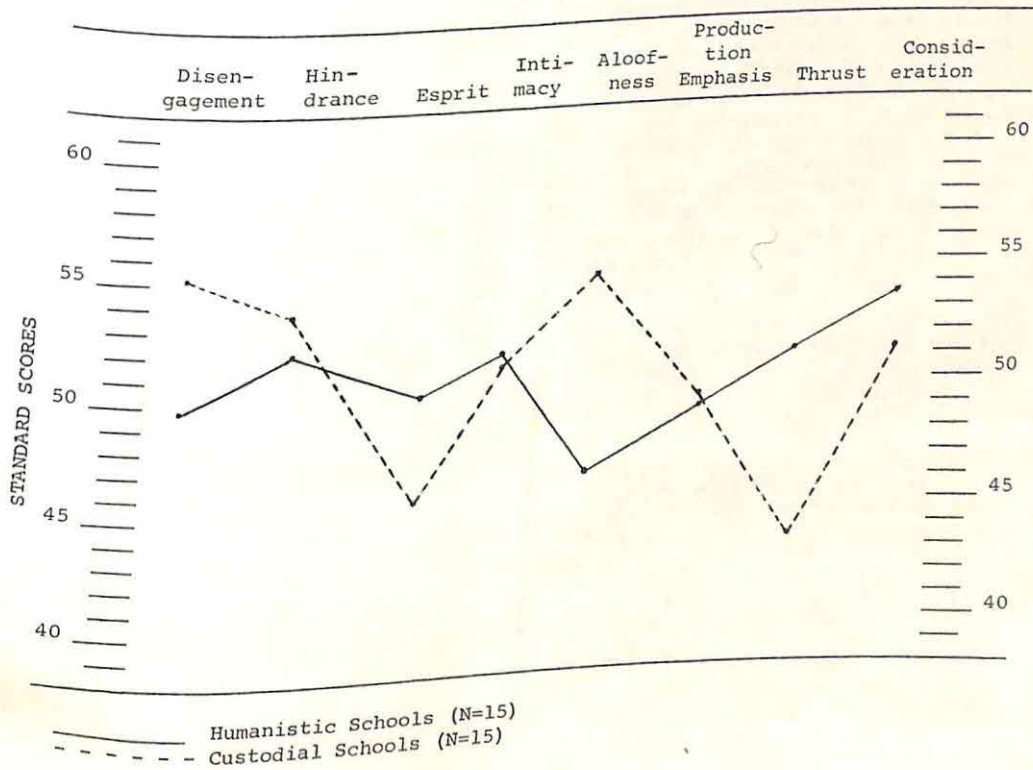
TABLE 1

SUMMARY DATA FOR HUMANISTIC AND CUSTODIAL ELEMENTARY SCHOOLS

SCHOOL CHARACTERISTIC	HUMANISTIC SCHOOLS (N=15)		CUSTODIAL SCHOOLS (N=15)		MEAN DIFFERENCE	F-RATIO
	MEAN	SD	MEAN	SD		
Teacher Behavior						
Disengagement	49.93	5.32	55.13	3.66	-5.20	9.71*
Hindrance	51.93	3.78	53.60	3.73	-1.67	1.47
Esprit	50.26	6.55	45.46	4.05	+4.80	5.82*
Intimacy	51.66	4.10	51.00	2.07	+0.66	0.32
Principal Behavior						
Aloofness	46.46	5.54	54.73	2.86	-8.27	26.35*
Production Emphasis	49.00	3.62	49.53	4.27	-0.53	0.14
Thrust	51.13	4.20	43.33	4.16	+7.80	26.02*
Consideration	53.66	4.46	51.26	2.31	+2.40	3.41
"Climate"						
Openness	51.47	13.27	33.67	8.79	+17.80	18.77*

* $p < .01$

indicates some important differences between the climate characteristics of the two types of schools. Analysis of variance computations yielded significant F ratios for differences between means of humanistic and custodial schools on Disengagement ($F=9.71$, $p < .01$), Esprit ($F=5.82$, $p < .01$), Aloofness ($F=26.35$, $p < .01$), and Thrust ($F=26.02$, $p < .01$). The degree of Intimacy and Production Emphasis was relatively the same in both types of schools. Teachers in humanistic schools experienced slightly less Hindrance and described their principals as more considerate than those in custodial schools. However, the differences between the means were not significant at the .05 level for either of these two dimensions ($F=1.47$ and $F=3.41$ respectively). These data are summarized in Table 1.

Furthermore, as hypothesized, elementary schools with a humanistic pupil control orientation were significantly more open than those with a custodial pupil control orientation ($F=18.77$, $p < .01$). The relevant data are also summarized in Table 1.

Although the present analysis focused on a contrast of schools in the sample with extreme pupil control orientation scores (upper third versus lower third), it is instructive to note that when coefficients of correlation were also computed, using data from all forty-five schools in the sample, parallel relationships emerged. Average pupil control ideology scores of schools correlated significantly with mean scores on Disengagement ($r=.40$, $p < .01$), Esprit ($r=-.49$, $p \leq .01$), Aloofness ($r=.67$, $p < .01$), Thrust ($r=-.60$, $p < .01$), and Openness ($r=-.61$, $p < .01$). Correlations between PCI and Hindrance ($r=.23$,

$p > .05$), Intimacy ($r=-.18$, $p > .05$), Production Emphasis ($r=.11$, $p > .05$), and Consideration ($r=-.25$, $p > .05$) were all not statistically significant (recall the higher the PCI score the less humanistic the school).

SUMMARY AND DISCUSSION

Humanistic schools were found to be different from custodial schools in several important ways. In addition to the basic contrast in orientations toward student control as measured by PCI scores, humanistic schools were more likely than custodial schools to have: (1) teachers who work well together, that is, pull together with respect to the teaching-learning task; (2) high morale and satisfied teachers, satisfaction growing out of a sense of task accomplishment and fulfillment of social needs; (3) principals who deal with teachers in an informal, face-to-face situation rather than "go by the book"; (4) principals who do not supervise closely but instead attempt to motivate through personal example; and (5) an atmosphere marked by openness, acceptance, and authenticity in teacher-principal interactions.

The data suggested that authenticity and openness in organizational behavior seem highly compatible with a humanistic pupil control orientation and incompatible with a custodial orientation. If interactions among teachers and between teachers and principals are authentic in humanistic schools, then it seems reasonable to hypothesize that authenticity will also tend to pervade teacher-pupil interactions. A humanistic pupil control orientation would appear

to facilitate, and be facilitated by, authentic interaction between teachers and pupils.

The importance of the concept of openness in the organizational climate of schools has been discussed in detail by Halpin and Croft; they posed the interesting query that perhaps ". . . climate profiles may actually constitute a better criteria of a school's effectiveness than many measures that already have entered the field of educational administration and now masquerade as criteria" (8:82-83). If the openness of the school climate provides one valid criterion of school effectiveness, then elementary schools with a humanistic pupil control ideology would appear to be significantly more effective, at least in terms of expressive or social emotional development, than those with a custodial orientation.

Moreover, to the extent that an elementary school attempts to communicate values as well as to communicate knowledge and develop skills, a humanistic pupil control ideology seems highly functional. A positive and strong commitment of students to the school seems required to effectively communicate values (5). It also appears unlikely that such commitment can be effectively attained in the custodial school: in fact, the custodial atmosphere in the school is more likely to produce alienation of students rather than commitment.

Although humanism and custodialism are descriptive terms assigned to contrasting pupil control orientations in elementary schools, it is difficult not to idealize the former through contrast with the latter. However, a word of caution seems in order. It is one thing to describe humanistic schools in a general way, but it is quite a different matter to find teachers equipped with demonstrably sound psychological and sociological theories necessary for the effective service of a humanistic approach. There are no simple approaches in changing the climate or atmosphere of schools. For example, recent research findings suggest that the pupil control ideology of beginning teachers becomes significantly more custodial as they become socialized by the teacher subculture, a subculture described by the vast majority of new teachers as one in which good control and good teaching were equated (9, 10, 11). More research is necessary to explore various strategies for changing the atmosphere of schools. For example, the study of the conflicts and adaptations of humanistic teachers attempting to teach in custodial schools and of custodial teachers working in humanistic schools might supply some useful clues in developing such a strategy.

In brief, the significance of pupil control orientation, as an important aspect of the organizational life of elementary schools, was underscored by the findings of this study. The concepts of custodialism and humanism provided a useful means for identifying schools with important differences in patterns of social interaction. If statements concerning orientation correspond relatively well with behavior, then the pupil control orientation of a school may provide another important step in identifying the "social climate" of the school.

FOOTNOTES

1. This research was supported in part by a grant from the Oklahoma State University Research Foundation.

2. For a complete discussion of the development of the PCI Form, see Donald J. Willower, Terry L. Eidell, and Wayne K. Hoy, The School and Pupil Control Ideology, Penn State Studies Monograph No. 24, University Park, Pennsylvania, 1967.

REFERENCES

1. Andrews, John, "School Organizational Climate: Some Validity Studies," Canadian Educational Digest, 5:317-34, December 1965.

2. Brown, Robert John, Organizational Climate of Elementary Schools, Educational Research and Development Council of the Twin Cities Metropolitan Area, Monograph Number 2, Minneapolis, Minnesota, 1955.

3. Carlson, Richard O., "Environmental Constraints and Organizational Consequences: The Public School and Its Clients," in Griffiths, Daniel E. (ed.), Behavioral Science and Educational Administration, University of Chicago Press, Chicago, Illinois, 1964, pp. 262-76.

4. Coleman, James S., The Adolescent Society, Free Press, New York, 1961.

5. Etzioni, Amitai, A Comparative Analysis of Complex Organizations, The Free Press, New York, 1961.

6. Gordon, C. Wayne, The Social System of the High School, Free Press, Glencoe, Illinois, 1957.

7. Halpin, Andrew W.; Croft, Don B., "The Organizational Climate of Schools," Administrators Notebook 11:7, March 1963.

8. Halpin, Andrew W.; Croft, Don B., The Organizational Climate of Schools, Midwest Administration Center, Chicago, Illinois, 1963.

9. Hoy, Wayne K., "Organizational Socialization: The Student Teacher and Pupil Control Ideology," Journal of Educational Research, 61:153-55, December 1967.

10. Hoy, Wayne K., "The Influence of Experience on the Beginning Teacher," The School Review, 76:312-23, September 1968.

11. Hoy, Wayne K., "Pupil Control Ideology and Organizational Socialization: A Further Examination of the Influence of Experience on the Beginning Teacher," The School Review, 77:257-65, September 1969.

12. Waller, Willard, The Sociology of Teaching, John Wiley and Sons, New York, 1967.

13. Willower, Donald J.; Eidell, Terry L.; Hoy, Wayne K., The School and Pupil Control Ideology, Penn State Studies Monograph Number 24, University Park, Pennsylvania, 1967.

14. Willower, Donald J.; Jones, Ronald G., "When Pupil Control Becomes an Institutional Theme," Phi Delta Kappan, 45:107-09, November 1963.

THE PRINCIPLE OF CONGRUITY

AS A PREDICTOR OF MEANING[1]

ROBERT B. KANE
Purdue University

WILLIAM B. RUDOLPH
Iowa State University

ABSTRACT

The ability of a congruity model to predict composite sign meaning as defined by responses to a semantic differential (SD) questionnaire was examined. The composite signs, component signs, and Ss were associated with the field of education. In most instances, obtained measures of factor scores were systematically lower than predicted measures. However, the addition of a constant c, such that $-0.3 \leq c \leq -0.2$, to the predicted measures generally removes this difference. Obtained and predicted factor scores were correlated to indicate their relationship independent of a systematic error. After accounting for the reliability of SD factor scores the correlations indicate that the congruity model does predict meanings of composite signs from meanings of component signs.

A SEMANTIC differential (SD) is a device which may be used to determine the connotative meaning of signs such as teaching or children. The meaning of a sign is defined as its point in Euclidean n-space with coordinates a_1, a_2,..., a_n. Each dimension of the meaning space for a sign is determined by a factor analysis of Ss' responses to a set of 7-point scales each defined by a pair of bipolar adjectives such as good-bad or hard-soft. Thus two signs having different meanings will be associated with different points in the meaning space.

What happens when two or more signs are present together? One might expect the meaning of teaching to interact with the meaning of children to educe the meaning of teaching children. Osgood, Suci and Tannenbaum (4: 199-216) developed a model to predict the meaning of a composite sign such as teaching children from the measured meanings of the component signs teaching and children. Their model is:

$$d_c = d_1 + \frac{d_1}{d_1 + d_2} (d_1) + \frac{d_2}{d_1 + d_2} (d_2)$$

where d is the deviation from neutrality on SD scales (i. e., the location on a 7-point scale from -3 to +3), c refers to the composite sign, and 1 and 2 refer to the first and second component signs respectively. The model is to be applied separately to each dimension of the meaning space. Osgood, Suci, and Tannenbaum (4:275-284) cite evidence to support the predictive power of their model. They report that:

(1) obtained factor scores for composite signs are consistently within the limits set by the factor scores of the components; (2) obtained factor scores deviated from the predicted scores on the average only by amounts attributable to unreliability except for factor I, the evaluative factor; (3) obtained and predicted factor scores exhibit a high positive correlation. They concluded that semantic effects follow the expectations from the congruity principle quite closely for the average meaning of composite signs.

This study was designed to determine whether or not the principle of congruity predicts composite sign meaning with component signs, composite signs, and Ss from elementary education.

METHODS

The Ss were seventy-one seniors majoring in elementary education at Purdue University and enrolled in their professional semester. Fourteen bipolar adjective scales were chosen by searching the literature for SD scales which consistently exhibited high and relatively pure loadings across a variety of signs judged by different kinds of Ss. A SD consisting of five component signs and four composite signs each to be rated on the fourteen scales was presented to each S. The component signs were mathematics, social studies, science, language arts, and teaching children. The composite signs were teaching children mathematics, teaching children social

TABLE 1

SD SCALES ASSOCIATED WITH EACH FACTOR

Factor I	Factor II	Factor III
happy-sad	heavy-light	fast-slow
good-bad	hard-soft	hot-cold
heavenly-hellish	difficult-easy	
positive-negative		
optimistic-pessimistic		

studies, teaching children science, and teaching children language arts. The order of sign and scale presentation was randomized as was the order of adjectives within scales. In-class time was used to administer the questionnaire; every S completed every item. Principal components factor analysis with rotation to Kaiser's varimax criterion (2,3) revealed that three factors accounted for 0.50 to 0.75 of the variance across scales among the nine signs. Table 1 lists the scales whose loadings were greater than 0.3 on their respective factors for at least seven out of nine signs for factors I and II, and for at least six out of nine signs for factor III. The remaining four scales were discarded since they were confounded across factors.

FINDINGS AND ANALYSIS

Scores for each S across the nine signs were calculated by computing mean scores for the set of SD scales within each factor. Predicted scores for each of the four composite signs were computed using the congruity model. Mean obtained and predicted scores over Ss are presented in Table 2.

A Z test for correlated data was used in comparing obtained and predicted means of composite concepts for factor I because the variances among scores for composite signs on factor I were not homogeneus. These means were significantly different ($\alpha < 0.01$). Since homogeneity of variance obtained among scores for composite signs on factors II and III, t tests for correlated data were used to analyze scores for these factors. Six out of eight differences were significant at $\alpha < 0.05$. The alpha level for each difference is displayed in Table 2.

It appears that the predictive power of the con-

TABLE 3

OBTAINED MEAN COMPONENT SIGN SCORES OVER Ss

	Obtained		
	Factor I	Factor II	Factor III
Language Arts	1.70	.47	.63
Mathematics	1.30	.77	.95
Science	1.67	.68	.31
Social Studies	1.23	.35	.34
Teaching Children	2.25	.13	.69

gruity model is somewhat stronger with factor II scores than with scores for factors I and III. In fact, the differences between obtained and predicted scores for factors I and III are significant at the 0.01 level in all but one case. Moreover, the predicted scores are consistently higher than the obtained scores for factors I and III. If a constant of about -0.3 were introduced into the prediction formula the differences between predicted and obtained scores for factors I and III would virtually disappear. The insertion of a constant of -0.3 would decrease the predictive ability of the formula in only one case among the factor II scores.

To obtain a different measure of the predictive validity for the congruity formula, mean component sign scores over Ss for each factor were calculated. These scores are displayed in Table 3.

Predicted means for the composite signs were computed by substituting the mean scores for the component signs into the congruity formula. Table 4 includes these predictions together with the obtained means for the composite signs.

Using t-tests for correlated data six of the differences between predicted and obtained mean scores are significant at the 0.01 level. The alpha level for each difference is displayed in Table 4. The pattern of differences between obtained and predicted scores when the predicted scores are generated from mean scores from component signs is quite similar to the

TABLE 2

MEAN FACTOR SCORES FOR FOUR COMPOSITE SIGNS OVER Ss

	Factor I			Factor II			Factor III		
	Obtained	Predicted		Obtained	Predicted		Obtained	Predicted	
Teaching Children Language Arts	1.81	2.10	($\alpha < .01$)	.67	.55	($\alpha < .40$)	.39	.69	($\alpha < .01$)
Teaching Children Mathematics	1.60	2.04	($\alpha < .01$)	.78	.99	($\alpha < .10$)	.32	.68	($\alpha < .01$)
Teaching Children Science	1.85	2.12	($\alpha < .01$)	.61	.85	($\alpha < .05$)	.58	.91	($\alpha < .01$)
Teaching Children Social Studies	1.60	1.99	($\alpha < .01$)	.18	.49	($\alpha < .02$)	.40	.66	($\alpha < .02$)

TABLE 4

PREDICTED AND OBTAINED MEANS FOR COMPOSITE SIGNS

| | Factor I | | | Factor II | | | Factor III | | |
	Obtained	Predicted		Obtained	Predicted		Obtained	Predicted	
Teaching Children Language Arts	1. 81	2 01	$(\alpha < .01)$	. 67	. 40	$(\alpha < .01)$	. 39	. 66	$(\alpha < .01)$
Teaching Children Mathematics	1. 60	1. 90	$(\alpha < .01)$	. 78	. 68	$(\alpha < .40)$	. 32	. 84	$(\alpha < .01)$
Teaching Children Science	1. 85	2. 00	$(\alpha < .05)$	. 61	. 59	$(\alpha < .50)$	. 58	. 57	$(\alpha < .50)$
Teaching Children Social Studies	1. 60	1. 89	$(\alpha < .01)$	. 18	. 29	$(\alpha < .40)$	. 40	. 57	$(\alpha < .05)$

pattern observable in Table 2. Prediction of factor II scores is better than prediction of factor I and III scores. In factors I and III, the predicted scores are higher than the obtained scores in all but one case. If the constant -0.3 were inserted in the congruity formula, predictions would be improved in six out of twelve cases. Predictions would be improved in eight out of twelve cases if the constant were -0.2. The data summarized in Tables 2-4 indicate that predictions of mean scores within factors based on the congruity formula may be improved by adding a constant.

Product-moment correlation coefficients between obtained and predicted scores over Ss were computed. These data, presented in Table 5, give an indication of the relationship between obtained and predicted scores which would remain invariant if a constant were added to each predicted score.

Test-immediate retest reliabilities of factor scores for seventh grade Ss were 0.84 for factor I, 0.72 for factor II, and 0.69 for factor III (1). While these coefficients might be expected to be somewhat higher for adult S's some of the correlations reported in Table 5 appear to be pushing their upper bound. All but the correlation for factor III under teaching children science are respectably high.

CONCLUSIONS AND RECOMMENDATIONS

The ability of a congruity model to predict composite sign meaning as defined by responses to a semantic differential questionnaire was examined. The component signs, composite signs, and Ss were all associated with teaching in the elementary school.

There were seventy-one Ss each enrolled in a professional semester for prospective elementary school teachers.

Two avenues of analysis were followed. First, a series of tests of differences between predicted and obtained measures of factor scores was completed. These data revealed a trend toward obtained measures being systematically lower than predicted measures. Thus, while the prediction model failed to "hit the mark," the adjustment of adding a constant, c, such that $-0.3 \leq c \leq -0.2$, to the predicted measures would have improved its marksmanship. Second, obtained and predicted factor scores were correlated to indicate their relationship independent of a systematic error such as the one described above. After accounting for the reliability of SD factor scores the correlations indicate that the congruity model does predict meanings of composite signs from meanings of component signs.

Additional research should confirm or refine the estimate that $-0.3 \leq c \leq -0.2$ is an optimum constant to use in revising the model for use with signs and Ss from the field of education.

FOOTNOTE

TABLE 5

CORRELATIONS BETWEEN OBTAINED AND PREDICTED COMPOSITE SIGNS OVER Ss

Teaching Children	Factor I	Factor II	Factor III
Language Arts	. 676	. 785	. 593
Mathematics	. 550	. 587	. 749
Science	. 608	. 615	. 372
Social Studies	. 505	. 565	. 519

REFERENCES

1. DiVesta, F. J.; DiVesta, Dick W., "The Test-retest Reliability of Children's Ratings on The Semantic Differential," Educational and Psychological Measurement, 26:605-616, 1966.

2. Kaiser, H. F., "The Varimax Criterion for Analytic Rotation in Factor Analysis," Psychometrika, 23:187-200, 1958.

3. Kaiser, H. F., "The Application of Electronic Computers to Factor Analysis," Educational and Psychological Measurement, 20:141-151, 1960.

4. Osgood, C. E.; Suci, G. J.; Tannenbaum, P. H., The Measurement of Meaning, University of Illinois Press, Urbana, 1957.

READING GROUPS AS PSYCHOLOGICAL GROUPS

PAT MCGINLEY[1] AND HUGH MCGINLEY
University of Wyoming, Laramie

ABSTRACT

Six classes of first-grade children were given a sociometric question asking them with which classmates they would prefer to work. The children were classified by their reading groups, and their in-group and out-group choices were analyzed. Each class was naturally trichotomized into three reading groups; top, middle, and lower. The reading groups were the only enduring groups of the classes. It was found that the lower reading group members chose fewer than expected children from their own groups ($p < .025$) and more than expected children from the top reading groups ($p < .001$). Members of the middle reading groups made fewer than expected choices from the lower reading groups ($p < .01$) and more than expected choices from the top reading groups ($p < .001$). The top reading group members made fewer than expected choices from the lower reading groups ($p < .001$), fewer than expected choices from the middle reading groups ($p < .06$), and chose within their own groups more than expected ($p < .001$). The results were discussed in terms of group cohesiveness and possible group effects upon learning.

CHILDREN can have, within the school framework, many and varied group experiences. Generally, school groups have been studied in terms of the criterion used for grouping and the resulting effects of the grouping on learning, and in terms of a group's psychological properties and their effects on individual behavior.

The educational researcher generally studies grouping procedures to determine their effects on learning. A number of the studies in grouping have involved reading groups; these studies usually use the skill of reading as a dependent variable and the grouping criteria as the independent variables. For example, the educational researcher has been concerned with the effect ability grouping has on reading development (18). Within this context, the definition of a group is a collection of individuals with some degree of homogeneity. The focal point of this type of research is on the verb form of the word group which is to arrange or to form.

The psychologist who studies groups is usually concerned with consistent and persistent behavior patterns which emerge within an interacting group. Behavior patterns which consistently emerge in groups and which persist through the life of groups are commonly referred to as being properties of the group. Some of the more common group properties that have been delineated are cohesiveness, conformity, leadership, and status. Within this context, the definition of a group is a collection of interacting individuals. As a result of the interaction, each member is changed by his group membership, and each member would probably undergo a change as a result of changes in the group (6). Of these group properties, Deutsch and Krauss (7) view cohesiveness as one of the most significant. Bonner (4) not only feels that cohesiveness is a basic group property, but feels that without at least minimal attraction among group members a group could not exist at all. In general, the group property of cohesiveness can be considered to be one of the fundamental dimensions of interpersonal attraction. Once a group is formed through interaction and interpersonal attraction, other group properties develop and group consequences emerge that effect each individual within the group. Lott and Lott define cohesiveness as: "That group property which is inferred from the number and strength of mutual positive attitudes among the members of a group" (13: 408).

Although the school class has been studied as a psychological group in order to describe its social structure (5, 10), instructional groups such as reading groups have usually been studied in terms of how their formation criteria affect reading development e.g., (15). However, these instructional groups have not been studied as social structures with behavior

patterns which may be affecting reading development.

The purpose of this study was to determine if reading groups develop into groups in the psychological sense. Since cohesiveness is recognized as a significant factor in the development and continuation of groups, a cohesiveness measure was used to determine whether or not reading groups develop into psychological groups. If a homogeneous collection of individuals in a reading group show positive attitudes toward each other by choosing each other for an activity, in preference to individuals in other reading groups, then the reading group may be postulated to be cohesive, and hence a psychological group. If individuals in the reading group choose individuals from other reading groups in preference to the members of their own, then it may be postulated that the reading group is not cohesive and therefore, not a psychological group.

METHOD

Subjects

A sociometric test was administered to six first-grade classes from two elementary schools located in Lexington, Kentucky. The schools were located in similar socioeconomic areas. One school was in an older middle–class residential area and the other in a newer middle–class suburban residential area. The classes were chosen by the school principals to meet the following criteria: a self–contained classroom; experienced teachers (suggested by the principals); average children who had little mobility in and out of school; relatively stable reading groups; and classes with the same number of reading groups. The class sizes ranged from twenty–three to twenty–nine with a total of 160 first graders.

In each class the children had been divided into three reading groups by the ability criterion. Group 1 was the high ability group, group 2 was of average ability, and group 3 was the low ability group. Each class had special names for the reading groups, which ranged in size from five to thirteen.

Out of the total number of children, one boy was a repeater, one boy had severly retarded speech, and one girl was mentally retarded (the teachers helped give the answers for the last two children). All but two of the children had been in their present classes for at least one semester. Several children were absent the first day of testing but were seen later in the week.

First graders were selected for this study for three reasons: (1) the classes are usually self–contained in the first grade; (2) the reading group is usually the only instructional group (in other grades math groups and other special activity groups are formed); and (3) first graders have had relatively little other formal group experiences, e.g. Brownies, Cub Scouts, etc.

Procedure

Each of the children was individually asked the following sociometric question:
If (teacher's name) was going to have you work with some other children in this class, which three chil-

dren would you like to work with?

The authors chose a sociometric question which employed the concept of work in order to convey a school activity as opposed to a play situation. The word work was not thought to be leading as opposed to other words which relate to specific concepts such as read.

The sociometric question was administered near the end of the children's first year of school. Each child was seen individually, outside the classroom. The question was given in the afternoon so as not to interfere with the morning reading program. The teacher introduced the interviewer and told the children they were going to be asked several questions about the class. The children were instructed to meet with the interviewer one at a time. The interviewer chatted with each child for a few minutes in order to establish rapport. The main sociometric question was asked, then the child was asked why he (she) chose each of the children he (she) named. Each child was thanked and then asked to tell the next child to come out. Each child was seen for approximately 5 minutes.

In addition to the sociometric question for the children, the teachers were asked (1) for the names of the children in each reading group and if any of the children had been recently changed from one group to another, (2) if they divided the class for any other enduring group activity with which the children might identify, and (3) which five children she felt were best in reading, sports, math, or were the most popular children, and any other information about the children she cared to share with the investigator. The last question was asked to determine the teachers' attitudes toward the children, and to determine whether the teacher saw the children as they saw each other.

RESULTS

Expected Choice Patterns

The first question asked in analyzing the data was: "Do the choices follow the expected pattern?" The expected values were the chance number of choices which would be received by members of each group if each group member had an equally likely chance of being chosen. The children's three choices were arranged into sociograms and tallied for chi–square analyses. [2] Table 1 lists the obtained (f) number of choices and the expected (F) choice frequencies for the reading groups of each of the six classes (School A, classes 1 through 3; and School B, classes 1 through 3), the overall chi–square for each class, and the overall chi–square for the total of the six classes. Five of the chi–squares were significant at beyond the .001 level. The chi–square for the sixth class was significant at the .005 level. The overall chi–square was also significant at beyond the .001 level. The overall chi–squares computed separately for School A and School B showed the same results.

Intragroup Choices: Cohesiveness

The second question asked in analyzing the data was "Were the reading groups cohesive?" Table 2 shows the chi–squares for the three reading group levels. The overall chi–square for the top reading group indicates that these groups made more (p<.001)

TABLE 1

OVERALL CHI-SQUARES FOR THE SIX CLASSES[a]

Class	Group	f	F		$(f-F)^2/F$		
A1	G1	40	21	=	17.19		
	G2	16	24	=	2.67	x^2	23.53**
	G3	22	33	=	3.67		
A2	G1	68	39	=	21.56		
	G2	14	27	=	6.26	x^2	40.01**
	G3	5	21	=	12.19		
A3	G1	38	24	=	8.17		
	G2	31	33	=	0.12	x^2	13.09*
	G3	18	30	=	4.80		
B1	G1	50	27	=	19.59		
	G2	14	21	=	2.33	x^2	30.45**
	G3	14	30	=	8.53		
B2	G1	48	33	=	6.82		
	G2	16	21	=	1.19	x^2	14.68**
	G3	5	15	=	6.67		
B3	G1	43	24	=	15.04		
	G2	16	24	=	2.67	x^2	21.38**
	G3	22	33	=	3.67		
	Overall chi-square[b]					x^2	$\overline{143.14}$

[a]df 2
[b]For the overall chi-square df 12 and $p < .001$.
*$p < .005$.   **$p < .001$.

intragroup choices and less intergroup choices than expected. The overall chi-square for the lower reading groups indicates that these groups made fewer intragroup choices and more intergroup choices than expected. The overall chi-square for the middle reading groups was nonsignificant but in the negative direction ($p < .11$, sign test).

Intergroup Choices

The third question asked in analyzing the data was "In what direction were the intergroup choices made?" Table 3 shows the expected (F) and obtained (f) intergroup choices. The top reading groups made fewer intergroup choices in the middle group ($p < .06$) and in the lower group ($p < .001$) than expected. The middle groups made more intergroup choices in the top reading groups ($p < .001$) and fewer intergroup choices in the lower reading groups ($p < .01$) than expected. The lower reading groups made more intergroup choices than expected in the top reading groups ($p < .005$).

Summary of Choice Patterns

The total results of the above three questions are summarized by Figure 1. The choices of the lower reading groups' members (number 3) can be described as follows:

1. The members did not choose members within their own groups (S-). This effect was significant at the .025 level.

2. The members chose in the middle reading groups slightly less than expected (NS-).

3. The members chose in the top reading groups more than expected (S+). This effect was significant at the .001 level.

The choices of the middle reading groups' members (number 2) were as follows:

1. The members made fewer than expected choices in the lower reading groups (S-). This effect was significant at the .01 level.

2. The members chose within their own groups slightly less than expected (NS-).

3. The members chose in the top reading groups more than expected (S+). This effect was significant at the .001 level.

The choices of the top reading groups' members (number 1) were as follows:

1. The members did not choose in the lower reading groups (S-). This effect was significant at the .001 level.

2. The members did not choose in the middle reading groups (S-). This effect was significant at the .06 level.

3. The members chose within their own groups more than expected (S+). This effect was significant at the .001 level.

From the results, it was concluded that the top reading groups were cohesive and the lower reading groups were not. In the middle reading groups there was a strong trend away from cohesiveness in favor of intergroup choices, although the data were inconclusive for a decisive statement about cohesiveness. Since the middle groups also made significantly more choices than expected in the top reading groups there is evidence in support of the assumption that the middle reading groups were not cohesive. The Top reading groups, using the measure of cohesiveness, were the only psychological groups.

FIGURE 1

DIRECTIONS OF INTER-AND INTRAGROUP CHOICES
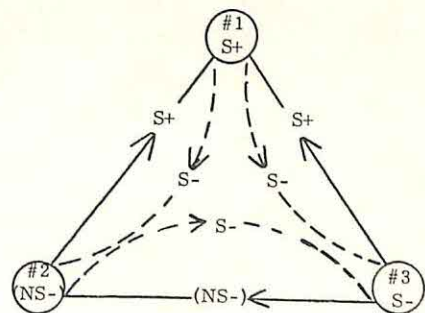
TABLE 2

CHI-SQUARES FOR INTRA-/INTER-GROUP CHOICES

Class	Group	Choice	f	F	$(f-F)^2/F$				
	G1	in	11	5.04	= 7.05				
		out	10	15.96	= 2.23	χ^2	9.28	df1	p <.005
A1	G2	in	3	6.72	= 2.06				
		out	21	17.82	= .80	χ^2	2.86	df1	NS
	G3	in	10	13.20	= .78				
		out	23	19.80	= .53	χ^2	1.31	df1	NS
	G1	in	32	16.71	= 13.99				
		out	7	22.29	= 10.49	χ^2	24.48	df1	p <.001
A2	G2	in	4	7.71	= 1.78				
		out	23	19.19	= .71	χ^2	2.49	df1	NS
	G3	in	2	4.50	= 1.39				
		out	19	16.50	= 0.38	χ^2	2.77	df1	NS
	G1	in	11	6.00	= 4.17				
		out	13	18.00	= 1.39	χ^2	5.56	df1	p <.025
A3	G2	in	11	11.79	= .53				
		out	22	21.21	= .29	χ^2	0.82	df1	NS
	G3	in	3	9.64	= 4.57				
		out	27	20.36	= 2.17	χ^2	6.74	df1	p <.01
	G1	in	22	8.64	= 20.66				
		out	5	18.36	= 9.72	χ^2	30.38	df1	p <.001
B1	G2	in	4	5.04	= .21				
		out	17	15.96	= .07	χ^2	00.28	df1	NS
	G3	in	7	10.80	= 1.34				
		out	23	19.20	= .75	χ^2	2.09	df1	NS
	G1	in	26	15.00	= 8.07				
		out	7	18.00	= 6.72	χ^2	14.79	df1	p <.001
B2	G2	in	7	5.73	= .28				
		out	14	15.27	= .11	χ^2	0.93	df1	NS
	G3	in	3	2.73	= .0256				
		out	12	12.27	= .0057	χ^2	.0313	df1	NS
	G1	in	9	6.46	= .998				
		out	15	17.54	= .367	χ^2	1.365	df1	NS
B3	G2	in	4	6.46	= .936				
		out	20	17.54	= .344	χ^2	1.28	df1	NS
	G3	in	8	12.69	= 1.73				
		out	25	20.31	= 1.08	χ^2	2.81	df1	NS

Overall chi-square –   Top Reading Groups χ^2 85.85   df6   p <.001
                         Middle Reading Groups χ^2 5.80   df6   NS
                         Lower Reading Groups χ^2 15.67   df6   p <.025

TABLE 3

INTERGROUP CHOICES

		G1 Selecting G2			G1 Selecting G3			G2 Selecting G1	
	f	F	$(f-F)^2/F$	f	F	$(f-F)^2/F$	f	F	$(f-F)^2/F$
A1	5	6.72	= .44	5	9.24	= 1.95	14	6.72	= 7.89
A2	5	12.54	= 4.53	2	9.75	= 6.16	22	12.54	= 7.14
A3	7	9.43	= .63	6	8.57	= .77	13	9.43	= 1.35
B1	4	7.56	= 1.68	1	10.80	= 8.89	11	7.56	= 1.56
B2	6	10.50	= 1.93	1	7.50	= 5.63	13	10.50	= .60
B3	4	7.39	= 1.55	11	10.15	= .07	17	7.39	= 14.47
		x^2	= 10.76 df5		x^2	= 23.33 df5		x^2	= 33.01 df5
			$p < .06$			$p < .001$			$p < .001$

		G2 Selecting G3			G3 Selecting G1			G3 Selecting G2	
	f	F	$(f-F)^2/F$	f	F	$(f-F)^2/F$	f	F	$(f-F)^2/F$
A1	7	0.56	= 1.20	15	9.24	= 3.59	8	10.56	= .62
A2	1	6.75	= 4.90	14	9.75	= 1.85	5	6.75	= .45
A3	9	11.78	= .66	14	8.57	= 3.44	13	11.79	= .12
B1	6	8.40	= .69	17	10.80	= 3.56	6	8.40	= .69
B2	1	4.77	= 2.98	9	7.50	= .30	3	4.77	= .66
B3	3	10.15	= 5.04	17	10.16	= 4.61	8	10.16	= .45
		x^2	= 15.47 df5		x^2	= 17.35 df5		x^2	= 2.76 df5
			$p < .01$			$p < .005$			NS

Reasons for Making Choices

Table 4 lists some of the children's reasons for making their choices. The terms friend and like were the two reasons most often given for the children's choices. It was felt that the terms friend, like, and nice were used almost synonymously; for example, "I like him," "He is nice," and "He is my friend." Some children used each term for each choice suggesting that when asked why they made their choices they felt they should not use the same term for each child. The term work ranked third and referred to a skill such as reading, writing, arithmetic, or coloring. The term play, the fourth ranked term, referred to "He plays with me," "I play with him," or "We play together." The term like was attached to most of the work and play answers; for example, "I like to play with him," "I like his work," "He likes to play with me," or "We like to play together."

The category of other was used for choices such as: "The teacher thinks she (the chosen child) is smart," "He loves me," "She kisses me," "He helps me catch the girls," etc.

Questions for Teachers

The teachers were asked four questions. Question 1 asked if the reading groups had been stable during the year or if the children had been moved from one group to another. The six teachers said their reading groups had been relatively stable, particularly for the 2 months prior to the study. In general, the changes made in the reading groups were made in the month of January, when two to four children were moved between the top and middle groups.

Question 2 asked if there were any other enduring group activities within the classroom. The teachers said that the reading groups were the only group activity organized by them that lasted over a 1-week period.

Question 3 asked the teachers to give the names of the five children they felt were the best in reading, math, sports, and the most popular in the class.

Four teachers gave five choices. Two of them

TABLE 4

THE STATED REASONS FOR THE CHILDREN'S CHOICES

Reasons	Number Choosing	Reasons	Number Choosing
Friend	107	Helps me	25
Like	73	Good	12
Work	52	Cute-pretty	10
Play	48	Needs help	9
Nice	47	Neighbor	9
Don't know	39	Other	49
		TOTAL	480

gave the names of children whom they listed at the top of the list of children in the top reading group. One teacher only named one child and the sixth teacher did not wish to rank the children. Fifteen of the twenty-one children who were named were members of the top reading groups.

DISCUSSION

The main question of the study, "Are reading groups psychological groups, that is, do they have the group property of cohesiveness?" has a split answer. The top reading groups were cohesive, therefore they were psychological groups. The middle and lower groups were not cohesive, therefore they were not, by our definition, psychological groups. Two questions immediately come to mind when viewing these results: "Why were the middle and lower reading groups different from the top reading groups in their choice patterns?" "What is the meaning or relevance of these results to reading development?"

For a full answer as to why there was a difference between the top reading groups and the middle and lower reading groups, further research must be conducted. However, speculative explanations can be made by applying the findings of small group research; for example, the research concerned with success—reward as antecedents of liking may be pertinent.

The success—reward antecedents of liking can be used as possible explanations why the members of the top reading groups like (choose) each other, and why the members of the lower reading groups tend not to like each other but like the top reading groups' members. Lott and Lott (12) suggest that success—reward influences liking in several ways: as an attraction to successful persons, as an attraction to persons sharing a successful experience, as an attraction to persons present in a success—reward situation, and as an attraction to the source of the reward. The last influence, attraction to the source of the reward, may be a reciprocal success—reward situation between the teacher and her pupils. Reciprocal liking can influence the atmosphere of an interpersonal situation to the extent that interaction develops. Both atmosphere and interaction can influence the development of liking. Success—reward is based on both the classical conditioning and operant conditioning learning models.

Attraction to Successful Persons. Gilchrist (8) found that successful persons tend to be chosen by both successful and unsuccessful persons. Berkowitz (3) found that partners who were successful in the task situation were liked more than partners who were unsuccessful in the task situation.

The children in the top reading groups were successful persons because they were excelling in a skill (task) that is a basic learning skill, and that has great value in our society. The skill of reading is a basic learning skill because most other school learning is based on it. Children's advancement in school is based on achievement, and tests of achievement are usually printed and are therefore indirect measures of reading ability.

Since the children in the top reading groups were successful, they were chosen more by the other children, both those who were successful and those who were unsuccessful in reading.

Attraction to Person Sharing a Successful Experience. Shelley (19) found that interpersonal liking was greater in groups experiencing a group success than in groups experiencing a group failure. The members of the top reading groups were sharing in a successful situation and therefore, they liked each other. The members of the lower reading groups were in an unsuccessful situation and were not attracted to one another (to failures).

Attraction to Person Present in a Success—Reward Situation. Lott and Lott (11) found that children developed liking for individuals present when reward was received. The children did not need to share in the reward directly, they just needed to be present when a child was rewarded.

Since the members of the top reading groups were successful in reading, each child was probably rewarded when he participated in reading. The other children who were present when the individual child was rewarded would be liked more than the children not present (other reading groups' members).

Attraction to the Source of the Reward. Persons tend to like the person (or persons) who give rewards (20). In the classroom this can become a reciprocal success-reward situation between the teacher and the children in the top reading groups. The members of the top reading groups read well, and the teacher rewards them. The children were attracted to the teacher (the source of their reward). The children's liking for the teacher could be a reward to the teacher, and she could be attracted to the children in the top reading groups (the source of her reward).

The teacher, feeling she was successful because the children were doing well and because they liked her, may have created a more relaxed atmosphere in the reading groups. A relaxed atmosphere can influence the amount of interaction which, in turn, can influence the development of interpersonal liking.

Implications of Success—Reward for the Middle and Lower Reading Groups. If the preceding interpretations are plausible, then it can be speculated that success-reward values well decrease as ability decreases. This would explain why the middle reading groups had slightly fewer in-group choices than expected, and why the lower reading groups made few in-group choices. The middle groups were experiencing a degree of success-reward and were expressing some degree of liking for the teacher. The atmosphere might not have been quite as relaxed as for the top reading groups but might have been enough for some degree of interaction to occur, and for some liking to develop.

The lower reading groups were not succeeding in reading in relation to the rest of the class. They knew they were poor in reading; the teacher did not reward them enough; their peers treated them as unsuccessful readers; their parents might not have rewarded them and might have even, unwittingly and subtly, punished them. In general, the situation was negative and one to be avoided. The teacher was not rewarding these children, and they were not rewarding her. The atmosphere might have been so strained that interaction

was reduced. Liking was probably at a minimum because of these negative forces.

Status in the Reading Groups

Status influences liking as either status similarity or as status dissimilarity. People in high status positions tend to like people with status similar to themselves. People who feel they belong to a lower status group tend to orient upward, and choose the high status individuals (9).

The members of the top reading groups were high status individuals in the class (in a relational sense to the two lower levels). The top reading groups were cohesive since high status individuals tend to choose one another, possibly recognizing high status similarity as well as similarity in reading ability and interests. The members of the middle and lower reading groups, being lower status individuals in the class, tended to orient their choices to the upward status positions.

The status and success—reward factors can combine into a set of interdependent forces which influence the development of liking. Low status readers might have chosen high status readers in order to identify with them, or because high status persons could give rewards or were receiving rewards.

Ability Grouping

The development of liking has been interpreted primarily in terms of the dynamics of the success—reward relationship and the status factors between the children. There was another factor, a very basic one, that may have influenced the development of the success-reward, status and liking factors. This factor was the criterion used in grouping for reading ability. Consideration needs to be given to the initial effects of placing children into groups by the ability criterion, especially when one considers the work of Robert Rosenthal.

Rosenthal and Lawson (17) randomly assigned rats to research assistants for several learning tasks. Some researchers were told that their rats were smart, while other researchers were told their rats were dumb. The group of supposedly smart rats learned better than the group of supposedly dumb rats. Rosenthal concluded that the experimenter's expectancy toward the rats affected the rats learning.

More recently, Rosenthal and Jacobson (16) have studied teacher expectancy toward children's performance. The experimental group was a random sample of first graders. The teachers were told that these children were late developers, and would suddenly improve their work performance. Rosenthal found that after a 2-year period, the children in the experimental group had made greater gains than the children in the control group. He concluded that the teacher's expectancy for the children's improvement affected her attitudes which, in turn, influenced the children's learning.

When teachers group for reading by using the ability criterion, there is a strong indication that they expect the children in the top reading groups to learn faster and better than the children in the lower reading groups. That is, the connotation of ability grouping is parallel to Rosenthal's rat study in that the top reading groups are considered "smart" children, and the lower reading groups "dumb" children. The teachers' choices in this study, of the "best" children in class coming from the top reading groups is some evidence for this statement. Also, five of the six teachers gave negative indications of liking for the children in the lower reading groups.

Not only do the teachers expect the top reading groups to be better and the lower reading groups to be poorer or slower but parents reflect this attitude, and the peer groups express a similar attitude. The children respond to this expectancy and develop accordingly. Mann (14) asked fifth graders to describe themselves and got answers like: " I am in the low fifth grade, I am too dumb," and " I happened to be a little smarter than the rest." Mann questions the emotional impact of ability grouping on children. Similar evidence was found by Axline (1) and Bell (2) while working with retarded readers. Both investigators found that the concept-of-self as a successful person and reader was very poor for these children. When the children were helped to strengthen their concept—of—self as a successful individual, reading improved.

Teachers group children by ability and expect the children in the top reading groups to read better than the children in the lower reading groups. These expectations combine with the attitudes of liking, because the teachers, as well as the children, will tend to like successful better than unsuccessful children. The teacher's expectations and liking for the children in the top reading groups probably influences the rewards given, the development of intragroup liking, the reciprocal success-reward factor, the atmosphere and the interaction, all of which influence the development of liking.

In a sense, the teacher stacks the cards against the children who are in the slower range of development, because once the teacher looks at her class in terms of top, middle, and lower ability, she develops an expectancy about the children that influences the development of reading and the development of interpersonal liking.

IMPLICATIONS

Since high group cohesiveness has been shown to facilitate learning, and low group or lack of group cohesiveness has been shown to be related to decreased facilitation of learning or to inhibit learning, the direct implications of this study are: (1) ability grouping in reading facilitates the top reading groups; (2) ability grouping in reading inhibits the learning of reading in the lower reading groups, and (3) ability grouping in reading is either slightly facilitating or slightly inhibiting to the learning of reading in middle reading groups.

The antecedents of liking, success-reward, atmosphere, interaction, and status influence reading development. The antecedents of liking and group cohesiveness combined with the teacher's expectancy toward the children in the top and lower ability groups also may affect reading development. If further research shows

these statements to be reliable then the value of ability grouping for reading instruction is open to question.

FOOTNOTES

1. This paper is based upon the first author's doctoral dissertation at the University of Kentucky.

2. The three choices for each child were not independent as once a child made a selection the selected child could not be chosen again. The bias created by this lack of independence for within subjects acts to slightly de-emphasize the greater than expected choosing of individuals in a particular group.

REFERENCES

1. Axline, V., "Non-directive Therapy for Poor Readers," Journal of Consulting Psychology 11: 61-69, 1947.

2. Bell, John, "Emotional Factors in the Treatment of Reading Difficulties," Journal of Consulting Psychology, 9:3-125, 1945.

3. Berkowitz, L., "Group Standards, Cohesiveness, and Productivity," Human Relations, 7:509-519, 1954.

4. Bonner, H., Group Dynamics, Ronald Press, New York, 1959.

5. Bovard, E. W., "Interaction and Attraction to the Group," Human Relations, 9: 481-489, 1956.

6. Cartwright, D.; Zander, A., Group Dynamics, Harper and Row, New York, 1968.

7. Deutsch, M.; Krauss, R. M., Theories in Social Psychology, Basic Books, Inc. New York, 1966.

8. Gilchrist, J. C., "The Formation of Social Groups Under Conditions of Success and Failure," Journal of Abnormal and Social Psychology, 47:174-187, 1952.

9. Hurwitz, J.; Zander, A.; Hymovitch, B., "Some Effects of Power on Relations Among Group Members," in Cartwright and Zander (eds.), Group Dynamics, Row, Peterson, Evanston, Illinois, pp. 800-809, 1960.

10. Jennings, H., Sociometry In Group Relations, American Council on Education, Washington, D. C. 1959.

11. Lott, A. J.; Lott, B. E. "Group Cohesiveness and Individual Learning" Journal of Educational Psychology, 57: 61-73, 1966.

12. Lott, A. J.; Lott, B. E., "Group Cohesiveness as Interpersonal Attraction. A Review of Relationships with Antecedent and Consequent Variables," Psychological Bulletin, 64:259-309, 1965.

13. Lott, A. J.; Lott, B. E., "Group Cohesiveness, Communication Level and Conformity," Journal of Abnormal and Social Psychology, 62: 408-412, 1961.

14. Mann, M., "What does Ability Grouping do to the Self—concept," Childhood Education 8: 351-360, 1960.

15. Ramsey, W. Z., "An Experiment with Three Plans of Grouping," Challenge and Experiment of the International Reading Association, Scholastic Magazines, New York, 1962b.

16. Rosenthal, R.; Jacobson, L., Pygmallion in the Classroom. Holt, Rinehart and Winston, Inc., New York, 1968.

17. Rosenthal, R.; Lawson, R., "A Longitudinal Study of the Effects of Experimenter Bias on the Operant Learning of Laboratory Rats," Journal of Psychiatric Research, 2:61-72, 1964.

18. Russell, D.; Fey, H., "Research on Teaching Reading," in Gage, W. L., (ed.) Handbook of Research on Teaching, Rand McNally and Co., Chicago, Illinois, pp. 865-928, 1963.

19. Shelley, H. P., "Level of Aspiration Phenomena in Small Groups," Journal of Social Psychology, 40: 149-164, 1954.

20. Solomon, L., "The Influence of Some Type of Power Relationships and Game Strategies upon the Development of Interpersonal Trust," Journal of Abnormal and Social Psychology, 61:223-230, 1960.

EXPRESS FUNCTIONAL RELATIONSHIPS AMONG

DATA RATHER THAN ASSUME "INTERVALNESS"

KEITH A. McNEIL
FRANCIS J. KELLY
Southern Illinois University, Carbondale

ABSTRACT

The statistical assumption of interval data is presented and questioned from a number of different angles. Suggestions are given as to how data might be transformed to map the criterion data. The crucial point is that in the behavioral sciences we never see the construct, only the criterion measures that we assume are good measures of the construct. Therefore, we can never verify that the criterion is an interval measure of the construct. Researchers should spend time in finding a meaningful criterion and then using nonlinear transformations to isomorphically map the independent predictor variable(s) onto the dependent criterion variable. The last example demonstrates that a given measure cannot be considered to be either interval or non-interval. Whether a variable isomorphically maps a criterion is a function of the theoretical system from which the researcher is working. Therefore it is inappropriate to refer to a variable as innately "interval."

MOST INFLUENTIAL statistical tests emphasize the necessity of assuming that the criterion is an interval scale for most statistical tests (e.g., F and t). This assumption is said to be necessary because the tests add numbers, an arithmetic process, which supposedly yields nonsense when a scale is not interval. The purpose of this paper is to examine the notion of intervalness and to suggest transformation activities which might result in isomorphic corresponding scales.

A scale is said to be interval when the numbers relate monotonically and rectilinearly to a construct. For example, if we have three pieces of rope—4 feet long, 2 feet long, and 6 feet long—we can say that the rope difference between 2 and 4 is the same as between 4 and 6; furthermore, the arithmetic average of the three lengths of rope is 4 feet ($3 \times 4 = 12 = 2 + 4 + 6$). In all of these operations the results make sense in the "real" world.

In contrast, the numbers in an ordinal scale are monotonic but not necessarily linear in relation to the construct. We may ask a student to rank in order, on the basis of interest, five college courses. The student may order them in the following manner: English, Mathematics, Art, Philosophy, and Music. We can assign numbers to the order: 5, 4, 3, 2, and 1. These numbers represent monotonicity (in this case in descending order), but we really have no information regarding the rectilinearity of the numbers to the construct interest. English (scale value 5) is of more interest than Mathematics (scale value 4), but can we say English is one unit more interesting than Mathematics? Probably not. Likewise we cannot say Mathematics (scale value 4) plus Music (scale value 1) combined are as interesting as English (scale value 5). Essentially, when using ordinal scales we go from most to least (or least to most) but we do not know the magnitude on the construct between adjacent pairs. In the case of interest, the student may like English, Mathematics, and Art quite a lot (the differences in interest might be small) and really detest Philosophy and Music. Additivity of units is not meaningful with the ordinal scale, and it is nonsense to say an object rank of 5 is five times greater than an object rank of 1. Please note, the conventional definition of intervalness of a scale depends upon the scale's monotonic and rectilinear relation to some construct or object (1). Note that if there is a perfect linear relationship between two variables, then the two variables are measuring a construct in the same way. An isomorphism exists between the two scales, and we refer to this as the two variables being intervally related. The variables may or may not be interval measures of the constructs. Indeed it is not important if the variables are interval measures of the construct, as long as they are meaningful and useful measures of behavior.

In psychological research, most of our criterion (dependent) and predictor (independent) scales are assumed to be measuring some underlying construct, but we seldom (never?) know the "real" nature of the construct. For example, we have invented the construct intelligence and have developed (IQ) tests to measure it, but we do not have a scale we can roll out of the person's head to determine how monotonically rectilinear the measurement scale (IQ units) conforms to the construct intelligence. Intelligence tests, nevertheless, do a fair job in predicting some behaviors we call intelligent (e. g., school achievement). In view of the fact that some tests more or less predict behaviors they theoretically should, we might desire to investigate the relationship of a test (given numerical values) with the numbers assigned to the criterion rather than the construct. To illustrate this point let us examine the ficticious data provided in Table 1.

TABLE 1

PROBLEM-SOLVING SCORES (Y) AND INTELLIGENCE TEST SCORES (X)*

	Y	X
Individual 1	4	2
2	16	4
3	36	6
4	64	8
5	100	10
6	144	12

* Two sets of scores by six individuals. X = values the individual is assigned as a result of completing an intelligence test. Y = the values assigned to the individuals based on his performances on job problem-solving.

The X variable is intelligence test scores and the Y variable problem solving performance scores. In

FIGURE 1

INTELLIGENT PERFORMANCE AS MEASURED BY NUMBER OF PROBLEM SOLUTIONS ON A JOB
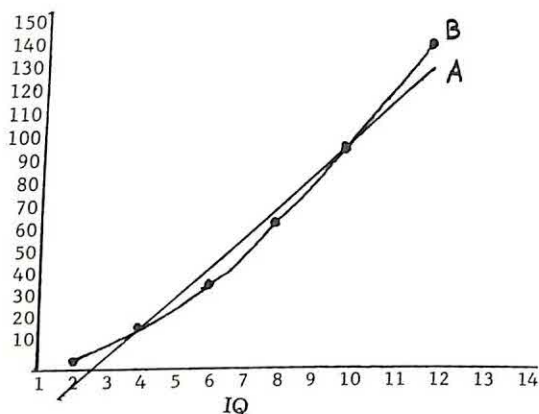
* Line A represents the line of best fit between X and Y using least squares solution $Y = a + bX$; $Y = -37.34 + 14X$. Line B represents the observed relationship between X and Y.

relation to Y scores, X is not an interval scale because a 2-unit increase on X from 2 to 4 yields a 12-unit increase on Y, whereas a 2-unit increase on X from 10 to 12 yields a 44-unit increase on Y. However, the X scale is an ordinal scale in relation to Y because monotonicity exists. One can use the least squares solution to solve for a and b in a regression equation $Y = a + bX$ to provide a line of best fit, which represents the degree of rectilinearity between X and Y.

The straight line of best fit does a fair job of representing the observed scores but tends to underestimate Y values of the extreme X scores and overestimate those mid-range scores. Table 2 shows the observed scores on Y, the predicted scores ($\tilde{Y}$), and a set of difference scores ($Y - \tilde{Y}$).

TABLE 2

OBSERVED SCORES ON Y, PREDICTED SCORES ($\tilde{Y}$) USING $\tilde{Y} = -37.34 + 14X$ AND DIFFERENCE SCORES ($Y - \tilde{Y}$)

Y	$\tilde{Y}$	$Y - \tilde{Y}$
4	-9.34	13.34
16	18.66	-2.66
36	46.66	-10.66
64	74.66	-10.66
100	102.66	-2.66
144	130.66	13.34

If we square each element in ($Y - \tilde{Y}$) and sum the squares we have the familiar error sums of squares within (ESS_W). This value is usually attributed to errors of measurement and lack of perfect association of the two measures to the underlying construct; but we also know that the assumption of a linear relationship in this case has been violated, therefore, some of the observed error must be due to adding and subtracting nonsense. We shall always be saddled with errors of measurement, but can we reduce the ESS by more closely approximating an interval relation between X and Y?

An inspection of Figure 1 suggests there may be a power relationship between scales X and Y. Suppose we assume X is ordinal to Y and therefore it would make sense to apply an area transformation to X. (A good educated guess would suggest squaring the elements of X.) Table 3 gives the Y, X, and X^2 scores.

It is obvious that the transformation results in an isomorphism of Y and X^2 scores. This relationship

TABLE 3

Y SCORES, X SCORES, AND X SCORES SQUARED

Y	X	X^2
4	2	4
16	4	16
36	6	36
64	8	64
100	10	100
144	12	144

can be expressed in a regression equation $Y = 0 + 1(X^2)$. Using this equation all predicted scores are numerically equivalent to the observed scores and thus $ESS_w = 0$. There is a linear relationship between Y and X^2 (the transformed scores). In this case the original ESS_w were attributed solely to the fact that the two sets of scores were not linearly related to each other. In reality we probably can't expect to observe such dramatic results due to transformations yet let us consider the possibilities that this ficticious bit of data provides.

When we measure and ascribe numbers to performance we often add the number of correct responses and use this value as the level of performance. If three or four of the items are extremely potent as they relate to the criterion, they will contribute to making a scale non-interval to the criterion. For example, if items 1-30 are equally easy and 31-34 are more difficult, then the difference between scores of 28-29 may be equivalent to the difference between 29-30 as they relate to the criterion (assuming these subjects missed items 31-34). However, the difference between 29-30 and 30-31 as they relate to the criterion will not be equal. The added unit increase (30 to 31) may yield a much larger increase of the Y scale than the unit increase from 29 to 30. Indeed, often we sum scores and really are not sure what the values mean to other scales or other observed behavior.

Since the advent of the computer, most research investigators seldom see a scatter plot of their data and often miss systematic departures from a rectilinear relationship among the data. The consequences may lead to an artificially large ESS due in part to departure from intervalness. Transformations may reduce this error.

In the preceding discussion as well as in the following illustrations, the reader should be aware of the focus of the numerical relationship. It is between the criterion and predictor(s). We may never know the quantitative nature of the underlying constructs, but if the criterion is meaningful in the context of one's theory, then a search for rescaling of the predictor(s) (and maybe even the criterion) to produce an approximate interval relationship seems legitimate, as long as the investigator realizes that the best transformation for one sample may not be generalizable to other individuals from the population. It is thus incumbent upon the investigator to replicate the transformation in successive samples from the population he is concerned about.

EXAMPLES OF NONLINEAR TRANSFORMATIONS

The following examples where a nonlinear transformation reduces error are provided to show a number of possible uses both with single and multiple predictors.

In all of the following examples we shall assume that the criterion is a meaningful measured behavior.

Example 1 (Figure 2)

Figure 2 represents a case where there is an interval relationship between the predictor and the criterion at the lower end of the predictor scale, but the high end of the scale does not meet the requirements of intervalness. A simple transformation on that end of the scale will create an interval X scale in relation to the particular criterion. It is in-

FIGURE 2

A SITUATION WHERE X IS AN INTERVAL MEASURE OF THE CRITERION ALONG PART OF THE SCALE
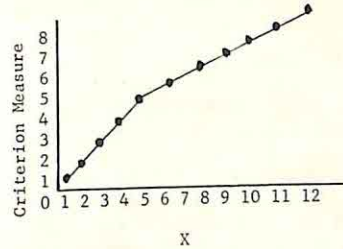
teresting to note that the linear transformation: $X' = 2.5 + .5X$ for those X values above 5 will suffice. Since the X values below the value of 5 do not need to be transformed, then with respect to the entire X scale, we say that we have made a nonlinear transformation.

FIGURE 3

A SITUATION WHERE TWO TRANSFORMATIONS ARE NEEDED
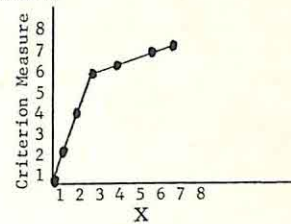
Example 2 (Figure 3)

In Figure 3, two transformations of the X variable will be necessary to create an interval relationship between X and the criterion. One transformation is necessary on X scores between 0 and 3, and another on X scores above 3. Both of these transformations will be linear transformations (within the range of interest). The only problem is to find the weighting coefficients which will do the trick. Some transformations are easy enough to spot without much calculation. Others are quite intricate and demand a very precise calculation. The multiple linear regression procedure can be very useful in this rescaling process (3).

An easy method to empirically find the necessary transformation is to find the weighting coefficients associated with the straight line. Let us refer back to the problem represented in Figure 2 to see how this can be done. We need to have two vectors, one which allows the Y intercept to manifest itself, and one which allows the slope of the line to manifest itself. The regression model which would accomplish this purpose is as follows:

$$Y_1 = a_0 U + a_1 X_1 + E_1$$

$$\begin{bmatrix} 5.5 \\ 6. \\ 6.5 \\ 7 \end{bmatrix} = 2.5 \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} + .5 \begin{bmatrix} 6 \\ 7 \\ 8 \\ 9 \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}$$

Since these weighting coefficients perfectly map the X values onto the Y values (all values in E_1 are equal to zero), we know that there is an interval relationship between X and Y. Now if we desire to make X an interval scale with respect to the criterion, all we have to do is simply apply these weighting coefficients to the X values and produce new X scores (X'):

$$X' = .5X + 2.5$$

Note that this is the same transformation that we performed for the data in Figure 2.

Now the problem in Figure 3 is a little more difficult since two transformations are necessary. The logic and procedure that is presented above is still applicable; all that is necessary is to construct vectors which will allow the slopes and Y intercepts of the two lines to manifest themselves.

$$Y_1 = a_1 U_1 + a_2 X_2 + a_3 U_2 + a_4 X_3 + E_2$$

$$\begin{bmatrix} 0 \\ 2 \\ 4 \\ 6 \\ 6\ 1/4 \\ 6\ 3/4 \\ 7 \end{bmatrix} = 0 \begin{bmatrix} 1 \\ 1 \\ 1 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} + 2 \begin{bmatrix} 0 \\ 1 \\ 2 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} + 5\tfrac{1}{4} \begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} + \tfrac{1}{4} \begin{bmatrix} 0 \\ 0 \\ 0 \\ 3 \\ 4 \\ 6 \\ 7 \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}$$

Where:

U_1 = 1 if the score on X is less than 3, zero otherwise;

X_2 = score on X if the value is less than 3, zero otherwise;

U_2 = 1 if the score on X is 3 or greater, zero otherwise;

X_3 = score on X if the value is 3 or greater, zero otherwise;

E_2 = error in prediction $Y_1 - \widetilde{Y}_1$; and a_1, a_2, a_3 and a_4 are partial regression weights calculated to minimize the error in prediction.

We could have solved both lines separately, but it seems that a simultaneous solution is more elegant, especially if a number of these lines must be found. Note that the X value of 3 is used in the determination of only one of the lines, although which line is used is entirely arbitrary as the data conform to both lines. The two transformations that are necessary are:

if X < 3, X' = 0 + 2X
if X ≥ 3, X' = 5 1/4 + 1/4 X

We will then have an interval mapping of X' on Y.

Some readers may question the appropriateness of such a transformation. Indeed we are uniquely fitting lines to a sample of data. Whether or not this is "the" transformation can to some extent be determined by randomly drawing another sample of data from the same population and applying the same transformation to that second set. If the correlation between the transformed X scores and the criterion is close to 1.0, then the transformation can be considered appropriate and X' is an interval scale with respect to the criterion under consideration.

Example 3 (Table 4)

The purpose of this example is to establish the non-necessity of the traditional notion of the assumption of interval scales. It appears to us that, with respect to the several goals of research—predictability, parsimony, replicability, validity generalization, control, and understanding—the assumption of an interval scale is really only necessary for the last named goal.

We shall now investigate a situation wherein both scales are clearly only ordinal measures. Scales H and I in Table 4 are both ordinal with respect to the underlying construct of length. For example, the 1-inch difference between lines A and B results in a difference of 12 units on the H scale, whereas the 1-inch difference between lines B and C results in a difference of 20 units on the H scale. Nevertheless, the correlation between scales H and I will yield an r value of 1.0. The transformation necessary to go from H to I is: I = 2 x H. This perfect correlation tells us that there is a 1-to-1 monotonic and rectilinear relationship between scales H and I. The point is that the correlation coefficient can be computed and the validity ascertained and tested for significance while one is fully aware of the fact that neither of the two variables is an interval measure. As a result of the perfect correlation between these two ordinal variables, one has met some of the goals of research, particularly those of predictability and parsimony. The goals of replicability, validity generalization, and control have not been investigated, although the ordinal nature of the scales does not obviate the attainment of these goals.

One would be hard put to argue that the use of ordinal scales can assist one in reaching the goal of understanding, for as Hays points out:

> Although the numbers standing for ordinal measurements may be manipulated by arithmetic, the answer cannot necessarily be interpreted as a statement about the true magnitudes of objects, nor about the true amounts of some property. (2:71)

TABLE 4

SIX LINES AND THREE SYSTEMS OF MEASURING THESE LINES

Line	Length Using Inch-Scale Units	Scale H	Scale I
A	1	4	8
B	2	16	32
C	3	36	72
D	4	64	128
E	5	100	200
F	6	144	288

Example 4 (Figure 4)

We have one further way to attack the traditional notion of intervalness and that involves the notion that a scale is either interval or it is ordinal. We are aware that most researchers would argue that there are degrees of intervalness in that some scales come closer to being interval measures than do others. Figure 4 (a and b) presents a situation wherein a scale is interval in one situation (Figure 4a) and quite ordinal in another (Figure 4b). It is interesting to note that the criterion and the way it is measured are exactly the same in both situations. There

FIGURE 4

A STATE OF AFFAIRS WHEREIN TWO SCALES ARE RECTILINEARLY RELATED TO ONE ANOTHER AND ANOTHER STATE OF AFFAIRS WHEREIN THE SAME SCALES ARE NOT RECTILINEARLY RELATED TO ONE ANOTHER, UNTIL A NONLINEAR TRANSFORMATION IS PERFORMED
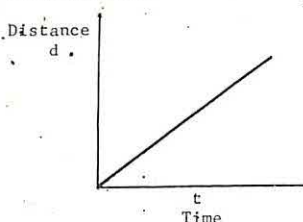
4a Observation of variables on the surface of the earth.
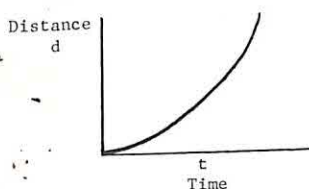
4b Observation of variables in vacuum moving towards the earth.
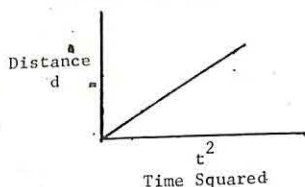
4c Observations of variables in vacuum moving towards the earth, after nonlinear transformation on time.
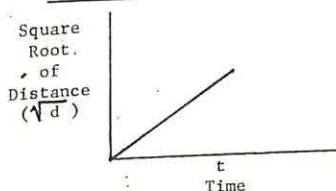
4d Observations of variables in vacuum moving towards the earth, after nonlinear transformation on distance.

is one minor difference and that is the system within which the variables are being related.

We usually think of distance (as measured in feet) and time (as measured in seconds) as interval measures. Indeed they are when we observe objects traveling along the surface of the earth (Figure 4a). A car traveling at a constant speed moves a very predictable distance in a given number of seconds, and for any given second, the car will travel a certain constant distance. Now look at Figure 4b, where we are concerned about studying the variables of distance and time in another instance—that of moving towards the earth. Now we do not observe an interval relationship between distance and time, although there appears to be a systematic relationship between the two variables. Which of the two variables is no longer interval? This seems to be a nonsensical question as it really doesn't matter, for we can transform t to perfectly map d, ($d = t^2$ as in Figure 4c), or we can transform d to perfectly map t, ($t = \sqrt{d}$, as in Figure 4d). What is important is that, as a result of the interval relationship between, say, t^2 and d, we can predict d if we know t; we can get this prediction in a relatively parsimonious fashion; we can check into the replicability of the relationship (indeed falling bodies in a vacuum will always follow the relationship); and we can also verify the validity generalization of the relationship (no matter what the mass, nor what the weight, all bodies will follow this relationship in a vacuum). In order for prediction to be accurate at various altitudes and on other planetary bodies, the concept of gravity must be entertained, (i.e., $d = 1/2 gt^2$). Furthermore, once we ascertain the above, we control the distance that an object will fall, and likewise we can control the amount of time we allow it to fall. We still may not understand why all this can be done, but at least we have made quite a few inroads into the phenomenon under consideration. And we have made all these discoveries with at least one ordinal scale!

The preceding comments pertain not only to correlational procedures, but also to analysis of variance procedures and any other procedure which is based upon the least squares solution. When the analysis of variance model is being used, the meaningful transformation will of necessity be on the criterion, unless there is some justification for ordinality of the independent variable.

The examples provided above have dealt with rescaling where monotonicity exists; however, there is no reason to assume that the two or more scales to be investigated must be monotonically related The literature relating test performance to anxiety measures often reveals a non-monotonic, non-rectilinear relation between the two measures. The transformation procedures outlined above can apply to the non-monotonic case as well, indeed the ∩ shaped function usually observed can easily be reflected by the equation: $\hat{Y} = a + b_1 X + b_2 X^2$ where the X variable represents anxiety and X^2 is the square of the elements in vector X. Such an equation will yield a positive value for weight b_1 and a negative value for b_2.

SUMMARY

Firstly, when one has what is known to be an ordinal scale, one may still have valuable information.

A nonlinear transformation may make the data more meaningful, or the criterion to be predicted may also be an ordinal scale, and the ordinality might just be of the same nature.

Secondly, even when the data are non-monotonic, transformations still may represent the relationship between scales such that the arithmetic processes are appropriate.

In essence, the authors believe we should forget about the "Holy Interval Scale" which exists in the minds of men and attempt to express the functional relationships among our data such that we reduce error due to adding and subtracting nonsense.

REFERENCES

1. Guilford, Joy P., Psychometric Methods, Mc-Graw Hill, New York, 1954, 605 pp.

2. Hays, William L., Statistics for Psychologists, Holt, Rinehart, and Winston, New York, 1965, 719 pp.

3. Kelly, Francis J.; Beggs, Donald L.; McNeil, Keith A.; Eichelberger, Tony; Lyon, Judy, Research Design in the Behavioral Sciences: Multiple Regression Approach, Southern Illinois University Press, Carbondale, Illinois, 1969, 353 pp.

BOOK REVIEWS

Robert E. Clasen

book review editor

THE TEACHING OF YOUNG CHILDREN: SOME APPLICATION OF PIAGET'S LEARNING THEORY

Brearley, Molly, Editor (New York: Schrochen Books, Inc., 1970), 191 pp.

A prerequisite to reading this short treatise on the teaching of young children is the reader's understanding of (a) Piaget's learning theory and (b) the British Infant School. The former is the educational foundation and the latter is the educational setting—yet, neither is given enough description, documentation, or explanation to be of real value to the uninformed reader.

Seven educators at the Froebel Educational Institute in England have compiled this book "to more precisely determine the goals of education for young children and how these goals may best be met." Except for the introduction written by Molly Brearley, Principal of the Institute, no credit is given for individual chapters which deal with: science, art, literature, movement, mathematics, music, morality, psychological standpoint, and teachers and children. Nor are professional biographies or even areas of professional interest listed. The reader is left with only a vague notion of the expertise underlying the presented points of view. This becomes of special importance when one looks beyond Piaget to the authors' "applications" of Piaget's learning theory. It is here that the book should have its greatest impact—yet, it is here that it has its greatest shortcoming.

The strong point of the book is Piaget and the authors' reaffirmation of the best we know about children and how they learn. (Many other experts on children and learning are referred to as frequently as Piaget. This raises the question of whether they should have received equal billing or if Piaget is a more marketable name.) Their thesis is "that the main value of the school is surely the fostering and developing of mental life enabling children to experience more fully and consciously all that life has to offer." Their plea to train teachers, by putting all the emphasis on understanding what one is doing and looking on all "aids" as means to clearly envisaged ends and as highly personal in use, hits at the very heart of current issues in early childhood education programs.

The chapter on science emphasizes scientific thinking rather than the too often teacher perceived role of "pounding in" the content of science. In the art chapter there is a good discussion of the stages leading to expression of representations through drawing, printing, modeling, and construction which aims at, but doesn't directly hit, the need for art as expression rather than a daily art product.

Under literature, the wide variety of literary forms are examined for the ways they can lay foundations for highly personal experiences as well as serving to extend language development. Movement is explained as a means of expressing and communicating and this chapter looks at the patterns children can self-select to show themselves as unique individuals.

The main function of the school in relation to mathematical development is described as providing children with the environment and helping to bring about a match between their personal psychological learning structures and the logical structures of mathematical knowledge. Ordering, relating, and measuring are emphasized.

Hearing and making sound get equal attention in the chapter on Music which stresses the need for an environment which allows children to discover and experiment for themselves with musical things and processes.

"Morality: Values and Reasons" deals with the concept of self and its relations with others, laying the foundation

INTERACTIONS OF ATTITUDES AND ASSOCIATIVE

INTERFERENCE IN CLASSROOM LEARNING[1]

WILLIAM L. MIKULAS
University of West Florida

ABSTRACT

Ss were tested on their knowledge, attitude, and prior experience toward one of two topics. Then they were assigned a reading or given a lecture on the relevant topic and retested after 2 days and 6 weeks, with half of the Ss given feedback after the second test. Analyses determined to what extent the following variables affected changes in attitudes and factual knowledge: topic, mode of instruction, feedback, prior attitude, prior factual knowledge, prior experience, and evaluation of the lecture or readings.

THE PROMINENT explanation of human forgetting of verbal material is associative interference theory (4) in which the major retention loss is due to competition from alternative responses at the time of recall. Most of the relevant studies, however, have utilized meaningless material such as nonsense syllables. Recent studies involving learning meaningful material (1, 2, 7) have failed to demonstrate simple interference effects, although in some cases meaningfulness may have been confounded with degree of original learning (8).

Classroom learning isn't as affectless as most verbal learning situations. Rather, the students often have very definite attitudes toward the material to be learned and the learning involves attitude changes as well as acquisition of content. Variables in the classroom that affect changes in student attitudes include the perceived credibility of the instructor, the personal involvement of the student in his position, the difference between the student's position and the position the instructor advocates, and how one-sided the instructor's message is perceived (5, 10).

An important educational objective is to provide the student with feedback concerning the accuracy and rate of his progress. Feedback may have any combination of the following effects: it can strengthen the student's learning as when he finds out he was right, it can result in motivational changes such as in the goals the student sets (6), it can change the direction the student is moving in his learning, and it can be a new learning experience or rehearsal of previous learning.

In the present study college students learned, by lecture or readings, about a topic on which they had definite attitudes and ideas. By assessing their attitudes and factual knowledge before the new learning, immediately after, and 6 weeks after, changes in attitudes and factual beliefs were investigated and compared with predictions from models of attitude change and verbal learning. Also, half of the students were provided feedback immediately after the second test. Most important, it was possible to determine interactions between the variables of verbal learning, attitude change, and feedback.

METHOD

Subjects

The Ss were the students of ten sections of an introductory psychology course with about twenty-five students per section. Eight of the sections comprised the eight possible experimental groups resulting from all combinations of the following three variables: Topic (hypnosis or therapy), Mode of learning (lecture or readings), and Feedback (present or absent). The other two sections comprised two control groups. As much counterbalancing as possible was done relative to the time the class met, the instructor's teaching method, the sex of the instructor, and course content.

Sequence of Events

All experimental Ss first took a pretest (test 1) on either hypnosis or therapy. Two days later they received

new learning in the appropriate subject area by one of two modes of instruction, lecture or readings. Two days and six weeks after the new learning the Ss were given retests, tests 2 and 3, over the material of test 1. Half the Ss received feedback over the correct answers to test 2 immediately after that test.

The two control sections had test 1 on one of the two topics, a lecture on the opposite topic, and tests 2 and 3 on the same topic as test 1.

INDEPENDENT VARIABLES

Topic. The content of the new learning was either hypnosis or therapy. The new learning on hypnosis, based on Hypnosis in Perspective (9), discussed basic phenomena of hypnosis and stressed the inherent dangers of being hypnotized, particularly by an unskilled hypnotist. "New Ways in Psychotherapy" (3), was the basis for the new learning on therapy which argued for "behavior therapy" as opposed to psychoanalysis.

Mode. The mode of the new learning was either by readings or lecture with both covering the same material. In the analyses comparisons were made between reading groups and lecture groups (Mode 2 analysis) or among reading groups, lecture groups, and groups that missed the new learning (Mode 3 analysis).

E assigned the readings and gave the lectures with the instructors of the classes telling Ss that the new learning was a course requirement. Ss were told before the new learning that the pretest emphasized certain misconceptions and central ideas which they would be retested on in 2 days with a test similar to the first.

Feedback. Following the new learning and test 2 on the new learning, half the Ss received feedback, a handout to read in class which identified the correct answers to test 2. E then answered any questions the Ss had.

DEPENDENT VARIABLES

Test 1. This pretest was composed of three parts:

(1) One of two short situational questions to assess the Ss' attitudes toward the particular topic: If the person you married became "mentally ill," would you send him to psychoanalysis? Would you allow yourself to be hypnotized by a doctor or dentist for medical purposes, or by an entertainer in a nightclub?

(2) Ten "factual" statements on which the Ss marked their degree of belief on 5-point scales: sure it is true, think it is true, don't know, think it is false, sure it is false.

(3) A biographical question asking Ss to list all previous psychology courses they had had plus any experience or training related to the previous questions.

Factual items for the tests were pretested on non-psychologists and an advanced psychology course. The final items were chosen such that there would probably not be any statements for which the Ss would be in general accord with the position of the new learning.

All three tests were prepared in two forms, A and B, which differed only in the wording of some of the statements, the situation question, and the order of the statements. The two forms were constructed to be equivalent, an aim suggested by the results of the present study. For although the tests were found to be highly reliable in terms of the effects of the main variables, the form of test 1 did not have a significant effect on either the attitude or factual scores of test 1. And the particular sequence of test forms between tests 1, 2, and 3 did not affect changes in attitude or factual scores between the tests.

On tests 1 and 2 a random half of each class was given form A and the other half form B. On test 3 each S had the opposite form from his test 2, or the opposite of test 1 if he missed test 2.

Test 2. Test 2 began with the same attitude question and ten factual statements as test 1. In addition Ss evaluated on 5-point scales (1 = very good, 2 = good, 3 = neutral, 4 = bad, 5 = very bad) the (a) quality, (b) credibility, and (c) bias (one-sidedness) of the readings or lecture. Or the S indicated that he was unable to do the readings or missed the lectures. To increase the validity of the Ss' answers about missing the new learning the E strongly emphasized that it didn't matter if they missed it but it did matter that they answer truthfully, and that their instructors would never see their answers.

For each topic the bias evaluations were grouped together in three approximately equal groups to form the variable Bias classification. For therapy the upper third contained scores 1-2; middle third, score 3; and bottom third, scores 4-5. For hypnosis the upper third contained score 1; middle third, score 2; and bottom third, scores 3-5.

Test 3. Test 3 consisted of the attitude question and the ten factual statements.

Scoring of Tests

Attitude scores. The following gives the scale values for the different responses to the attitude question on hypnosis.

score 1:   S would allow himself to be hypnotized by anyone.

score 2:   S would allow anyone to hypnotize him but only with certain restrictions.

score 3:   S would only allow a "qualified" person (e. g., doctor or dentist) to hypnotize him.

score 4:   S would only allow a qualified person to hypnotize him under certain conditions.

score 5:   S would never allow himself to be hypnotized.

A similar scale was used to assign scores to the answers to the therapy attitude question:

score 1: S would use psychoanalysis;

score 2: S would use psychoanalysis but with some restrictions;

score 3: S would use psychoanalysis but only in conjunction with other procedures;

score 4: S would use psychoanalysis after trying other approaches;

score 5: S would not use psychoanalysis.

The attitude scores were divided by topic into three approximately equal groups to form the variable Attitude classification. For therapy the top third contained score 1; middle third, score 2; and bottom third, scores 3-5. For hypnosis the top third contained scores 1-2; middle third, score 3; and bottom third, scores 4-5.

Factual scores. In addition to scores given the statements by Ss, each response was marked "right" or "wrong" relative to the new learning. Thus, if S marked a statement 2 (think it is right) while the lecturer said it was false, this item would be marked wrong. Scores of 3 (don't know) were classified as "right" since the emphasis in this study was on items about which the Ss had misconceptions. In the later analyses of factual changes classifying a score of 3 as right is logically equivalent to not assigning it any label at all.

The variable Factual classification was constructed by dividing the factual scores, separated by topic, into three approximately equal groups. For therapy the top third contained scores 0-3; middle third, scores 4-5; and bottom third, scores 6-10. For hypnosis the top third contained scores 0-5; middle third, scores 6-8; and bottom third, scores 9-10.

Experience scores. In response to the question about prior experience in the area, S was scored 1 if he had no prior experience; 2 if he had had some general courses related to the topic or had done some reading in the area; and 3 if he had had personal experiences in the area, courses specifically dealing with the topic, or some personal involvement in the issues.

RESULTS

A series of fifty-eight, independent, two-way analyses of variance were done using the following classification variables:

Topic: hypnosis or therapy; [2]

Mode 2:   lecture or reading;

Mode 3:   lecture or reading or missed the new learning;

Feedback: received feedback or did not;

T1 form: form A or form B of test 1;

Attitude classification: the third of all test 1 attitude scores on a given topic in which a particular attitude score fell, i.e., top, middle, or bottom third;

Factual classification: the third of all test 1 factual scores on a given topic in which a particular factual score fell;

Experience: the score of the answer to the question about prior experience in the area;

Bias classification: the third of all evaluations of bias for a given topic in which a particular score fell;

Test sequence: the sequence of the forms of the 3 tests used for a particular S, e.g., ABA and BBA;

Section Number: the particular section S was in.

Throughout the results main effects and interactions listed as not significant are those that were not statistically significant at less than the .05 level Because of the number of statistical tests computed, it is possible that some of the apparently significant findings occurred by chance. However, the consistency of the results makes this improbable, at least for the main effects.

Test 1 Results: Attitude Scores and Factual Scores

Attitude scores. An analysis of variance on the scores of the answers to the attitude question showed Topic to have had a significant effect (F= 38. 60; df = 1,144; p< .0001) with the scores on the hypnosis tests (mean = 3. 02) higher than those on the therapy tests (mean = 1. 76). That is, Ss in the Hypnosis groups were hesitant about being hypnotized, but Ss in the Therapy groups were fairly sure they would send their spouses to psychoanalysis.

The following did not have a significant effect on the attitude scores:

T1 form
Section number X T1 form
Experience
Topic X Experience

Section number was by necessity significant since Topic was significant and each section had only one topic.

Factual scores. The factual scores are the number of factual items out of ten marked "correct" on the basis of the Topic material. As was true of the attitude scores, Topic had a significant effect on the factual scores (F = 73. 79; df = 1, 144; p< .0001) with the mean score on hypnosis tests (6. 93) higher than the mean for the therapy tests (4. 19). That is, Ss incorrectly marked more statements about therapy than about hypnosis, although it is improbable the tests were of equal difficulty. As with the attitude scores, the following did not significantly affect the factual scores:

T1 form
Section Number X T1 form
Experience
Topic X Experience

Test 2 Results: Evaluations of Quality, Credibility, and Bias

On test 2 all Ss evaluated the quality, credibility,

and bias (one-sidedness) of the new learning. Since most of the Ss marked the quality and credibility very good or good, regardless of Topic or Mode, these variables were not considered further.

The bias evaluations showed more dispersion with Topic significantly affecting the scores (F=72.67; df =1,95; p< .0001). The lectures and readings on therapy were considered by the Ss to be more biased than those on hypnosis. The following did not significantly bias scores:

    Mode 2
    Topic X Mode 2
    Attitude classification
    Factual classification
    Attitude classification X Factual classification
    Experience
    Topic X Experience

Changes Between Tests 1, 2, and 3

Factual changes. For each factual item the change in its score relative to the new learning was computed between test 1 and test 2 and between test 1 and test 3. Changes between test 2 and test 3 were not computed directly as this would violate the statistical assumption of independence for the analyses. If an S marked an item with a 2 (think it is right) on test 1 and marked the corresponding item with a 5 (sure it is false) on test 2, this would be a change of three units. If the new learning said the item was false it would be a change of +3; and if the new learning said it was true, it would be a change of -3.

Table 1 lists those variables which significantly affected factual changes between test 1 and test 2, while Table 2 lists the changes between tests 1 and 3.

The order of the attributes of the corresponding variables in Tables 1 and 2 is the same. For example, Topic has two attributes, therapy and hypnosis, and there was a greater factual change in the Therapy groups than the Hypnosis groups, both between test 1 and test 2 and between test 1 and test 3. Lecture produced a greater change than readings (Mode 2 analysis) and both of these produced a greater change than no new learning (Mode 3 analysis). Ss with a medium amount of experience showed the greatest change, Ss with the most experience changed least, and Ss with no experience were in between. The lower the S's factual score on test 1 (the fewer the number of

TABLE 1

VARIABLES AFFECTING FACTUAL CHANGES BETWEEN TESTS 1 AND 2

Variable	df	F	p <
Topic	1, 97	5.45	.025
Mode 3	2, 124	26.16	.0001
Mode 2	1, 97	9.74	.005
Experience	2, 92	3.41	.05
Factual classification	2, 124	5.09	.01

TABLE 2

VARIABLES AFFECTING FACTUAL CHANGES BETWEEN TEST 1 AND TEST 3

Variable	df	F	p <
Topic	1, 84	11.75	.001
Mode 3	2, 111	24.11	.0001
Mode 2	1, 84	9.97	.005
Experience	2, 68	3.20	.05
Factual classification	2, 109	8.94	.001
Feedback	1, 88	8.83	.005
Bias classification	2, 68	4.43	.025

correct responses), the greater the change. Those groups that received feedback after test 2 showed a greater change on test 3 than groups without feedback. Although Bias classification did not significantly affect factual changes between tests 1 and 2, it had a significant effect on the factual changes between tests 1 and 3. Ss who judged the new learning as very biased showed the greatest factual change on test 3, while Ss who judged it as the least biased showed the least change.

The following variables and interactions did not have a significant effect on the factual changes between test 1 and test 2:

    Bias classification
    Test sequence
    Attitude classification
    Topic X Mode 2
    Topic X Mode 3
    Topic X Factual classification
    Topic X Attitude classification
    Topic X Bias classification
    Topic X Experience
    Mode 2 X Attitude classification
    Mode 2 X Factual classification
    Mode 2 X Bias classification
    Experience X Bias classification
    Factual classification X Attitude classification

While these variables and interactions did not have a significant effect on the factual changes between test 1 and test 3:

    Test sequence
    Attitude classification
    Topic X Mode 2
    Topic X Mode 3
    Topic X Factual classification
    Topic X Attitude classification
    Topic X Bias classification
    Topic X Experience
    Topic X Feedback
    Mode 2 X Feedback
    Mode 2 X Factual classification
    Mode 2 X Attitude classification
    Mode 2 X Bias classification
    Mode 3 X Feedback
    Experience X Bias classification
    Factual classification X Attitude classification

TABLE 3

VARIABLES AFFECTING ATTITUDE CHANGES BE-
TWEEN TESTS 1 AND 2

Variable	df	F	p <
Topic	1, 124	5.34	.025
Mode 3	2, 124	5.12	.01
Attitude classification	2, 121	10.20	.0001
Mode 2 X Bias classification	2, 94	4.90	.01
Attitude classification X Factual classification	4, 121	2.87	.05

For two of the variables, Topic and Factual clas-
sification, the level of significance increased between
test 2 and test 3. For both variables the attribute
producing the biggest change (therapy, low factual
classification) showed an increase between test 2
and test 3. While the attribute with the smallest
change (hypnosis, high factual classification) show-
ed a decrease.

Attitude changes. For each S the amount of change
in attitude, positive or negative, relative to the new
learning (score 5) was computed between tests 1
and 2 and between tests 1 and 3. The variables hav-
ing a significant effect on changes in attitude are shown
in Tables 3 and 4.

As with factual changes, the order of the attributes
of the variables significantly affecting attitude change
is the same between tests 1 and 2 as between tests 1
and 3. In the case of Topic the Ss in the Therapy
groups showed a greater attitude change on test 2
than Ss in the Hypnosis groups. The Therapy groups
also showed a greater change on test 3. In the Mode
3 analysis (lecture vs readings vs missed the new
learning) the lecture groups showed more change
than the readings groups which showed more change
than the groups that missed the new learning. But
since the Mode 2 analysis (lecture vs readings) was
not significant, the effect of the Mode 3 analysis was
due to the lecture and readings groups having signif-
icantly more change than the group that missed the
new learning. In the case of Attitude classification,
the less the Ss' original attitudes agreed with the new
learning, the greater the attitude change. This rela-
tion of attitude classification to attitude change paral-

TABLE 4

VARIABLES AFFECTING ATTITUDE CHANGES BE-
TWEEN TEST 1 AND TEST 3

Variable	df	F	p ≤
Topic	1, 111	6.33	.025
Mode 3	2, 111	3.39	.05
Attitude classification	2, 107	8.00	.001
Mode 2 X Bias classification	2, 70	4.50	.025

lels the effect of Factual classification on factual
change.

The interaction between Mode 2 and Bias classi-
fication revealed that for Ss who perceived the new
learning as very biased or very unbiased the lecture
was more effective for changing attitudes. While for
Ss who perceived the new learning as being moderate-
ly biased, the reading was more effective. The signif-
icant interaction between Attitude classification and
Factual classification on test 2 is somewhat more
complex: In the case of low factual scores on test 1,
the attitude change is greatest for Ss with extreme
attitudes on test 1. In both cases this consisted of a
drifting toward the mean, i.e., Ss with low attitude
scores showed positive changes and those with high
attitude scores showed negative changes. This may
simply be due to simple regression toward the mean.
In the case of high factual scores on test 1, the atti-
tude change was the greatest for Ss with intermediate
attitude scores on test 1; the changes here being in a
positive direction. Ss intermediate in their factual
scores on test 1 were intermediate in performance to
the two cases above.

The following variables and interactions did not
have a significant effect on the attitude changes be-
tween test 1 and test 2:

Mode 2
Experience
Bias classification
Test sequence
Factual classification
Topic X Mode 2
Topic X Mode 3
Topic X Factual classification
Topic X Attitude classification
Topic X Bias classification
Mode 2 X Factual classification
Mode 2 X Attitude classification
Experience X Bias classification

The following variables and interactions did not
have a significant effect on the attitude changes be-
tween test 1 and test 3:

Mode 2
Feedback
Experience
Bias classification
Test sequence
Factual classification
Topic X Mode 2
Topic X Mode 3
Topic X Experience
Topic X Feedback
Topic X Factual classification
Topic X Attitude classification
Topic X Bias classification
Mode 2 X Feedback
Mode 2 X Factual classification
Mode 2 X Attitude classification
Mode 3 X Feedback
Experience X Bias classification
Attitude classification X Factual classification

Control group results

In addition to control data derived from analyses
such as with the variable Mode 3, two groups were

specifically included to test for halo effects of the E. For these groups, as described in the method section, the three tests were on one topic while the new learning was a lecture on the opposite topic. A series of t-tests determined if the amount of attitude or factual changes between tests 1 and 2 or between tests 1 and 3 was significantly different than zero. None of the tests yielded values significant at the .05 level.

DISCUSSION

Evidence for test reliability comes from the consistency of the F-ratios and order of the attributes between tests 2 and 3, which were 6 weeks apart. For example, lecture produced a bigger factual change than reading on test 2 with F = 9.74 and again lecture was more effective than reading on test 3 with F = 9.97. Inspection of Tables 1-4 shows this consistency to generally hold up for both factual and attitude scores.

Some of the results can most simply be explained that the greater the room for improvement the greater the change that will occur. The Therapy groups showed more factual change than the Hypnosis groups because the Therapy groups had significantly lower factual scores (number of " correct" items) on test 1. Similarly, analysis of Factual classification showed that across both topics the lower the Ss' factual scores on test 1, the greater the amount of factual change on tests 2 and 3.

A parallel situation is true of changes in attitude. Hypothesis groups had significantly higher attitude scores on test 1 than Therapy groups. Since the new learning argued for attitudes with high scores, the Therapy groups showed greater changes in attitude than the Hypnosis groups. Similarly the lower a S's attitude score on test 1 (according to Attitude classification) across both topics, the greater the amount of attitude change.

Change due to Mode, Experience, and Feedback

The superiority of lectures over readings for changing attitudes and factual beliefs may be because, particularly with controversial material, the lecturer can alter his presentations to fit each particular class and because lectures more easily elicit emotional responses. However, generalizations are limited because there was only one lecturer and two readings.

The effect of Experience was that the Ss with a medium amount of prior experience showed the most factual change, those with no prior experience the second most, and those with considerable experience the least. The small factual change with high experience Ss is probably because these Ss already knew the most and had the least room for change and / or were the most personally involved in their position and hence the most resistant to change.

One explanation for medium experience Ss showing more factual change than low experience Ss is that a small amount of experience (e.g., some popular reading) might give the S wrong information. If this happened, the medium experience Ss would have more room for improvement than the low experience Ss. Experience did not have a significant effect on test 1 scores. A more probable explanation is that medium experience Ss had more interest in the topic than low

experience Ss and hence learned more, but it can't be said which came first, interest or experience.

The significant effect of Feedback on factual changes was impressive because its manipulation was quite minimal and the effects weren't measured until after 6 weeks. Observation of Ss during the feedback session suggested they were quite interested in receiving feedback even though they knew they weren't being graded on the test.

Associative Interference and Attitudes

The 2-factor verbal-learning model of forgetting (4) predicts that over time pro-active inhibition (PI) will increase relative to retroactive inhibition (RI). Thus it might be expected that there would be a change in the factual scores from test 2 to 3 in the direction of test 1. That is, when S learns a set of factual material (new learning) that is contradictory to his prior learning (as measured by test 1), then when tested soon after the new learning (test 2), there is a change in performance in the direction of the new learning. However, when tested considerably later (test 3) there should be a regression toward his prior position (test 1).

The present experiment did not find such regression, except for one attribute of Topic and Factual classification. Generally there was no evidence for any regression between tests 2 and 3, and in the case of the most dominant attribute of Topic and Factual classification there was an increase rather than a regression. Also most theories of retention would predict more forgetting during the 6 weeks between test 2 and test 3.

A possible explanation for the failure of the data to fit the simple interference model is that the time intervals employed in this study were significantly longer than those of the usual verbal learning studies. It might be argued that all predicted effects from the interference model are over by the time of test 2. There is a report (7) of an increase in the PI of descriptive prose up to 7 days, but this was with non-controversial material.

A second explanation that might be concluded from Underwood and Ekstrand (11) and Mills and Winocur (8) is that the learning measured by test 1 was well learned under distributed practice and this didn't extinguish easily, decreasing the relative increase in PI over time. Such an explanation, based on studies of affectless material, may be partially true, but it appears necessary to take into account the Ss' attitudes.

The proposed explanation is that on test 2 the Ss were aware of the conflict between their previous ideas and those argued for in the new learning, and although they were willing to concede some facts, they had resistance about completely agreeing. This is suggested by remarks two Ss added to the bottom of their test forms: "Your lecture made me think. It made me question psychotherapy seriously, but I still find it hard to completely accept the behavioral method and reject the other." " According to your lecture many of my answers above are wrong, but I'm hesitant to give in that easily."

It is further proposed that 6 weeks later, at the

time of test 3, there is a reduction in resistance which more than offsets any forgetting. The reduction may be due to a dissipation of the motivational state that resulted from the conflict between the prior beliefs and the new learning and/or it could be due to a gradually built up confusion of the S about what he originally believed and the ideas of the new learning.

Consider Bias classification which was not significant on test 2 but was significant on test 3. On test 3 those Ss who had judged the new learning to be the most biased showed the biggest factual change; while those who had judged it to be the least biased showed the least factual change. The following set of relationships might hold: The greater the difference between S's test 1 scores and the new learning, the greater the room for change but also the greater the conflict. And the greater the conflict the greater the resistance to change and the more biased the S is apt to perceive the new learning. Thus on test 2 Ss who perceive the new learning as very biased have the greatest room for factual change, but this is offset by their also having the greatest resistance to change. Hence Bias classification does not have a significant effect on test 2 factual scores. But on test 3 when resistance has decreased, the Ss with the highest bias scores have the most room for change and in fact change the most.

The effect of Topic on factual changes increased in significance from the .025 level to the .001 level between tests 2 and 3. Topic had a significant effect on bias scores.

Any explanation for the results of the present study or what actually takes place in a classroom must take into account the interactions between the content and presentation of the new material and the attitude-belief structure of individual Ss.

FOOTNOTES

1. This study was conducted while the author was associated with the Center for Research on Learning and Teaching at the University of Michigan The author would like to thank Stanford C. Ericksen for his support and suggestions. Requests for reprints should be sent to the Faculty of Psychology, The University of West Florida, Pensacola, Florida, 32504.

2. The effects of the variable Topic are confounded with the fact that the hypnosis and therapy tests were not equated for difficulty. But for the purpose of this study Topic is used only to determine the relative performance on these particular tests without making inferences about how the populations respond to these topic materials. This is particularly so since Topic did not significantly interact with any other variable.

REFERENCES

1. Ausubel, D. P.; Stager, M.; Gaite, A. J. H., "Retroactive Facilitation in Meaningful Verbal Learning," Journal of Educational Psychology, 59: 250-255, 1968.

2. Ausubel, D. P.; Stager, M.; Gaite, A. J. H., "Proactive Effects in Meaningful Verbal Learning and Retention," Journal of Educational Psychology, 60: 59-64, 1969.

3. Eysenck, H. J., "New Ways in Psychotherapy," Psychology Today, 1: (no. 2) 39-47, 1967.

4. Keppel, G., "Retroactive and Proactive Inhibition," in Dixon, T. R.; Horton, D. L. (eds.) Verbal Behavior and General Behavior Theory, Prentice-Hall Englewood Cliffs, New Jersey, 1968.

5. Kiesler, C. A.; Collins, B. E.; Miller, N., Attitude Change: A Critical Analysis of Theoretical Approaches, John Wiley, New York, 1969.

6. Locke, E. A.; Cartledge, N.; Koeppel, J., "Motivational Effects of Knowledge of Results: A Goal Setting Phenomenon?" Psychological Bulletin, 70: 474-485, 1968.

7. Mills, J. A.; Sacks, S., "Proactive Inhibition of Descriptive Prose as a Function of the Length of Retention Interval," Journal of General Psychology, 76: 241-249, 1967.

8. Mills, J. A.; Winocur, G., "Retroactive Inhibition of Connected Discourse as a Function of Meaningfulness and Degree of Original Learning," Journal of Psychology, 71: 69-76, 1969.

9. Moss, C. S., Hypnosis in Perspective, MacMillan, New York, 1965.

10. Rosnow, R. L.; Robinson, E. J. (eds.), Experiments in Persuasion, Academic Press, New York, 1967.

11. Underwood, B. J.; Ekstrand, B. R., "An Analysis of Some Shortcomings in the Interference Theory of Forgetting," Psychological Review, 73: 540-549, 1966.

AN ANALYSIS OF TWO SOCIAL STUDIES

PROGRAMS AND FIRST GRADE ACHIEVEMENT

IN ECONOMICS

ROBERT F. SCHUCK
University of Pittsburgh

ROBERT F. DEROSIER
Hall High School, West Hartford, Connecticut

ABSTRACT

This study examined the influence of two widely used instructional programs on first-grade-level pupil achievement in economics. A total student population of 116 was studied. Six intact groups from four elementary schools in West Springfield, Massachusetts, were divided into three groups of two each. One group was assigned the Science Research Associates (SRA) materials for a full year, one group was assigned the use of the Follett materials for a full year, and the third group spent a semester using each of these programs. Data were gathered through the administration of the Spear's Test For Achievement In Economics in a pretest-posttest fashion. The data collected were submitted to both analysis of variance and covariance techniques. The results of this study indicate that pupils taught with a full year of SRA materials achieved significantly higher than pupils taught with a full year of Follett. When intelligence and pretest scores were held constant the combination SRA and Follett group scored significantly higher in pupil achievement than the Follett group alone. When teacher understanding of economics concepts was held constant, the combination group did not score significantly higher than the Follett group on pupil achievement.

WHILE a review of pertinent literature showed a number of studies concerned with teaching economics at the elementary level, empirical research, which has measured the effectiveness of instruction, has, until most recently, been virtually nonexistent. The 1960 volume of the Encyclopedia of Educational Research does not list a single study of economic education at the elementary-school level. The considerable disparity in the number of research programs devoted to the elementary grades, as compared to intermediate, secondary, and adult levels, is due in part to the requirements for testing first-grade pupils. For example, criterion instruments and research designs have to be substantially different in testing first-grade pupils as compared to testing adolescents.

Jefferds was among the first investigators attempting to measure empirically first-grade pupil achievement in the understanding of economics (2:41). His study attempted to measure a specific and widely used set of instructional materials — the Senesh materials — published by SRA, to determine if children being instructed by these new materials received any educational advantages over children taught by conventional methods.

Based upon his findings, Jefferds concluded that: (1) "children did no better on the criterion test with packaged instructional materials than with local materials"; and (2) "there were no significant differences in children's understanding of economic concepts regardless of how the instructional unit was presented."

Recognition of this lack of effective measuring instruments for economic education influenced the Spears study, conducted in Culver City, California (4: 89). Based upon the data collected, he concluded that: (1) under appropriate instructional programs, first-grade pupils can learn sophisticated economic concepts; (2) experimental Ss' learning of economics was significantly different from chance performance at the knowledge level; (3) experimental Ss' learning of economics was not significantly different from chance at the application level; and (4) the study indicated that the learning style of low socioeconomic

pupils places them at a disadvantage in the first-grade curriculum of Culver City, California.

While there is existing empirical evidence that first-grade pupils can comprehend economic concepts, there has been no attempt to differentiate among the increasing number of instructional programs at the elementary level. The rejection or adoption of instructional programs by school districts is therefore frequently based only upon the opinion of teachers and not upon empirical evidence.

PROBLEM

The basic purpose of this study was to compare the effectiveness of two widely used sets of social studies instructional materials in teaching economics to pupils at the first-grade level. The following null hypotheses were tested. (1) There will be no significant differences in achievement in economics for first-grade pupils among groups which are instructed:

1. with SRA materials for 2 semesters (Group A) and with Follett materials (Group B) for 2 semesters;

2. with SRA materials for 2 semesters (Group A) and with Follett materials for 1 semester and SRA materials the other semester (Group C);

3. with Follett materials for 2 semesters (Group B) and with Follett materials for 1 semester and SRA materials the other semester (Group C).

(2) There will be no significant differences in achievement in economics among pupils in different socioeconomic levels within each of the groups or among the groups.

SRA provides extensive instructional materials, largely prescriptive in design, which have as their objective teaching pupils basic social science understanding through the discovery process. The SRA materials consisted of the following items: (1) a textbook, Our Working World: Families at Work, 272 pp.; (2) a 198-page resource unit for teachers containing aids, suggestions, and activities for the fifty-five lessons in the textbook; (3) a 71-page picture workbook for the pupils; (4) a 49-page handbook for teachers which contained the written transcription of the records used to supplement each of the last twenty-seven lessons of the textbook. All of these materials were developed under the direction of Lawrence Senesh.

The Follett program has similar goals; however, it is less extensive, less prescriptive, and allows greater flexibility for the teacher. Follett materials contain the following: (1) two textbooks for the pupils, Billy's Friends, 143 pp. and Exploring with Friends, 168 pp. (The former text was used in the first semester's instruction; the latter in the second semester); (2) a 63-page teacher's guide for Billy's Friends and a 64-page teacher's guide for Exploring with Friends. Both guides contained objectives, teaching aids, and suggestions for a first-grade social studies program.

DESIGN

This study utilized six intact, first-grade classes from four elementary schools of West Springfield, Massachusetts. Pupils from the six classes were com-

bined into three groups, and each group was instructed with a different set of social studies instructional materials. The instrument used to measure the effectiveness of the three programs of instructional materials was Spears' Test for Achievement in Economics.

The scores of the 116 pupils in the study were analyzed according to the program with which they were instructed, and their socioeconomic status. Pupils were pretested in January 1969, and posttested in May 1969.

Pupil social position was determined by using Hollingshead's Two Factor Index of Social Position. This method utilizes the occupational and educational levels of the family's main wage earner as determinants of social position. A pupil information form was developed to secure this information. Scale values were found for the two factors, and these values were then statistically treated to provide an index of each pupil's social position. Pupils were classified into three socioeconomic levels: (1) level one represented pupils of the upper middle-class; (2) level two represented pupils of the middle-class; and (3) level three represented pupils of the lower middle-class.

Analysis of variance treatment was used for determining if significant differences in achievement in economics existed among the groups. Two-way analysis of variance treatment was used for determining if significant differences in achievement in economics, based on socioeconomic status, existed within or among the groups. Analysis of covariance treatment was used to secure statistical equalization on certain relevant variables which could have confounded the relationships under investigation.

DETERMINING TEST INSTRUMENTS

The instrument used to measure achievement of first-grade pupils in economics was developed by Sol Spears. The test consisted of twenty-six multiple-choice items.

In consideration of the maturity and attention span of first-grade pupils, this test takes approximately 30-33 minutes to administer. Due to the limited reading ability of first-grade pupils, the instructions for the pre- and posttest were tape-recorded in order to make them identical for all groups. Items for the instrument were obtained from a review of sources including textbooks, course outlines, statements of objectives, and questions from other tests.

A jury of economists who are subject matter experts in economic education —Dr. Norman Townshend-Zellner and Dr. John Lafky —examined the economic content aspects of the test, and endorsed it for validity. Only those items on which there had been complete agreement by both judges were included in the test. The reliability coefficient for Spears' Test for Achievement in Economics is .78, which was deemed acceptable for group testing purposes.

SELECTION OF INSTRUMENT MEASURING SOCIAL POSITION

Hollingshead's Two Factor Index of Social Position was chosen as the instrument for determining pupil social position (1:116). This instrument met the

need for an objective, easily applicable procedure to estimate the position individuals occupy in the status structure of our society.

According to Hollingshead, the following breakdown is meaningful for predicting the social class position of an individual:

Social Class	Range of Computed Scores
I	11-17
II	18-27
III	28-43
IV	44-60
V	61-77

Social class I represents the highest social position, and social class V the lowest.

For purposes of this study, an arbitrary delineation was made for pupil groupings. Level 1 was composed of pupils falling within social classes I and II. Level 2 was composed of pupils located within social class III. Level 3 was composed of pupils falling within social classes IV and V.

In order to determine the socioeconomic status of the Ss, a pupil information form was developed. All pupils were requested to take the form home for a parent (or guardian) to complete. To secure parent cooperation, a covering letter which briefly described the purpose of the study, and firmly assured the parent that no names would be used in the study, was sent to each pupil's home. All the forms were returned, and every parent supplied the desired information. The data obtained from the form were used to establish a scale score which was then weighted to give pupils a social class position.

PUPIL POPULATION AND TEACHER SELECTION

The Ss for this study were selected from the first grades of four elementary schools in West Springfield, Massachusetts, a suburb of 26,070 in 1965. The district has nine elementary schools with an enrollment of 3,097 pupils. There are eighteen first-grade classes in the district.

The six teachers who participated in the study were selected at random. All of the eighteen first-grade teachers had indicated a willingness to participate in the study. However, only fifteen met the following criteria: (1) completion of 3 or more years of successful teaching, and (2) a rating, by their administrators, of above average or excellent.

Course background in economics was very sparse for the six teachers. Two teachers had a 3-hour course more than 10 years before, two teachers had an in-service course in economics offered by Boston University in 1964-1965, and two teachers had never had any course work in economics. Thus, the mean hours of college course work completed in economics by the six teachers, was less than 3.

All elementary teachers in the system experienced a 30-hour in-service training program in the teaching of economics. The same economist taught all the teachers, and at the conclusion of the training program, all teachers were given the Teacher's

Economic Understanding Test. The results of this test provided data used to equate, statistically, the influence of any variance in teacher knowledge of economic materials.

Prior to the actual pupil testing time, the researcher met with each participating teacher to explain the purposes and procedures of the testing program. At that time it was made clear that only the effectiveness of the instructional materials, not teacher competence, was to be measured.

ADMINISTRATION OF TEST INSTRUMENT

January 16 and 17, 1969, were selected for the administration of the pretest to all Ss. The posttest was administered May 28 and 29, 1969. The teachers never saw the measuring instrument used in the study, nor were there any meetings between the teachers and investigators from January to May.

The investigators scored all of the tests in order to insure consistent scoring. All of the 116 tests were usable, and formed the basis from which data was obtained.

ANALYSIS OF THE DATA

Major Hypothesis

There are no significant differences in pupil achievement as measured by the Spears'test among pupils exposed to SRA materials for 2 semesters (Group A), pupils exposed to Follett materials for 2 semesters (Group B), and pupils exposed to SRA materials for 1 semester and Follett materials the other semester (Group C).

Findings

Using the raw scores pupils received on the Spears' test, an analysis of variance was applied to determine the F value of the difference. The data cited in Table 1 shows the test of significance with the Spears' test among the groups. The F value for the difference was determined to be 8.84, and the F table indicated that 7.37 (an approximate interpolated value) was needed to be significant at the .001 level of confidence. The null hypothesis of no significant differences among Groups A, B, and C was therefore, rejected.

TABLE 1

ANALYSIS OF VARIANCE FINDINGS ON POSTTEST ACHIEVEMENT SCORES GROUPED ACCORDING TO INSTRUCTIONAL MATERIALS

Source of Variation	df	Sum of Squares	Mean Sum of Squares	F Ratio
Instructional Materials	2	107.83	53.92	8.84 *
Error	113	688.96	6.10	

*Significant at the .001 level of confidence.

The Scheffe method was applied to the posttest scores of all the groups and the findings were:

1. Group A (SRA materials all year) scores were significantly greater than Group B (Follett materials all year) scores;

2. Group A scores were not significantly different from Group C (Follett materials 1 semester and SRA materials 1 semester) scores;

3. Group B scores were not significantly different from Group C scores.

The null hypothesis of no significant difference in achievement in economics between pupils of Group A and Group B was, therefore, rejected.

The null hypothesis of no significant difference in achievement in economics between pupils of Groups A and C failed to be rejected. The null hypothesis of no significant difference in achievement in economics between Groups B and C failed to be rejected.

At first glance, the strong findings favoring the SRA materials appear convincing. However, tests were conducted on relevant variables to determine whether or not the groups were truly matched. Therefore, analyses of variance were conducted for age in months, intelligence scores (IQ), socioeconomic status (SES), and pretest scores. These variables were chosen because much of educational literature is concerned with the relationship of chronological age, intelligence, and socioeconomic status, to learning ability. Pretest scores were used because they are also significant measures of the groups' equivalence at the start of the study.

Table 2 presents the means, standard deviations, and F ratios of the variables mentioned in the previous paragraph, grouped according to instructional materials (IM).

No significant differences based on age or on socioeconomic status were found among the groups. Significant differences were found among the groups for intelligence scores and pretest scores.

Again the Scheffe method was applied to the intelligence scores to determine which group was significantly different from another. The findings were:

1. Group A was not significantly different from Group B;

2. Group A was not significantly different from Group C;

3. Group B was significantly greater than Group C.

The Scheffe method was also applied to the pretest scores, and the findings were:

1. Group A was significantly greater than Group B;

2. Group A was significantly greater than Group C;

3. Group B was not significantly different from Group C.

Socioeconomic Status

Table 3 presents the number of Ss, mean scores, standard deviations, and F ratios with the variables — age (in months), intelligence scores, pretest scores, and posttest scores —grouped according to socioeconomic status.

No significant differences were found for the variables listed in Table 3, when pupils were grouped according to socioeconomic status (SES). An F ratio of 3.11 (an approximate interpolated value) at the .05 level, is needed for significance. The data cited in Table 3 show none of the F ratios to be significant. The null hypothesis of no significant differences in achievement among pupils grouped according to socioeconomic status failed to be rejected.

Possible Interaction Between Socioeconomic Status and Teaching Materials

Popham stated on interaction: "The general principle involved in interaction effects is the same in all analysis of variance models, when the research is testing for the existence of a relationship between the

TABLE 2

SUMMARY OF F VALUES FOR VARIABLES GROUPED ACCORDING TO INSTRUCTIONAL MATERIALS

	Group A (n=38)		Group B (n=40)		Group C (n=38)		F Ratio
	$\overline{X}$	sd	$\overline{X}$	sd	$\overline{X}$	sd	
Age	86.00	3.61	84.03	4.67	84.03	5.24	2.39*
IQ	106.03	12.07	109.21	9.22	101.69	10.03	4.69**
SES	2.45	0.76	2.40	0.87	2.58	0.68	0.55*
Pretest	11.74	2.40	10.43	2.07	9.89	2.56	6.21***

* Not significant.
** Significant at .05 level of confidence.
*** Significant at .01 level of confidence.

TABLE 3

SUMMARY OF F VALUES FOR VARIABLES GROUPED ACCORDING TO SOCIOECONOMIC STATUS

	Level I			Level II			Level III			F Ratio
	n	$\overline{X}$	sd	n	$\overline{X}$	sd	n	$\overline{X}$	sd	
Age	20	84.35	4.58	21	84.09	4.01	74	84.93	4.80	0.33*
IQ	20	110.60	8.22	19	104.58	9.29	71	104.69	11.61	2.51*
Pretest	20	10.70	2.34	21	11.29	2.53	75	10.51	2.46	0.83*
Posttest	20	11.60	3.00	21	12.42	2.04	75	11.42	2.67	1.20*

*Not significant at the .01 level of confidence.

dependent variable and another variable" (3:129). In accordance with Popham's observation, tests were performed, and F ratios were obtained to determine possible interaction of socioeconomic status and instructional materials with age, intelligence scores, pretest scores, and posttest scores.

Sex

No significant differences were found for the variables listed in Table 5 when pupils were grouped according to sex. To be significant at the .05 level, an F ratio of 3.11 (an approximate interpolated value) is needed for significance. The data cited in Table 5 show none of the F ratios to be significant.

DISCUSSION

Data resulting from this study indicate that the pupils of Group A who were instructed with the SRA materials for 2 semesters, scored significantly higher on the posttest of Spears' Test for Achievement in Economics than the pupils of Group B, who were instructed with Follett materials for 2 semesters. However, Group A did not score significantly higher on the achievement test when compared to pupils of Group C, who were instructed with Follett materials for 1 semester and SRA materials the other semester. Also, Group C did not score significantly higher than Group B on the posttest.

TABLE 4

SUMMARY OF F RATIOS FOR POSSIBLE INTERACTION BETWEEN SOCIOECONOMIC STATUS AND TEACHING MATERIALS

Variance	Sources of Variance	F Ratio
Age	SES x IM	0.65*
IQ	SES x IM	0.54*
Pretest	SES x IM	0.02*
Posttest	SES x IM	0.76*

* Not significant at .05 level of confidence.

No significant differences were found to exist among the groups on age, intelligence scores, socioeconomic status, pre-and posttest scores, when two-way analysis of variance was applied to pupils' scores, within and among the groups, for possible interactions between socioeconomic status and the instructional materials.

No significant differences were found to exist among the groups on age, intelligence scores, socioeconomic status, pre-and posttest scores, when the data were analyzed according to sex.

Treatment of Rival Hypotheses

As Table 2 indicates, significant differences did exist among groups on the variables of pupil intelligence and pretest scores. This being so, two rival hypotheses for the differences found among the groups on the posttest could be formulated, i. e. :

1. The differences noted were not due to the instructional program employed, but rather to the differences in pupil intelligence;

2. The differences noted among groups were not due to the instructional program employed, but rather to the fact that the groups were not equal at the start of the investigation;

and an additional rival hypothesis also presents itself, i. e. :

3. The differences among groups noted was not due to the instructional program employed, but rather to the fact that the teachers varied in their understanding of the substantive material taught (economics).

Each of the rival hypotheses was investigated, using analysis of covariance to hold constant the effect of the variable in question.

Rival Hypothesis I. The differences noted were due to the differences in intelligence of pupils, not the instructional program employed. Table 2 indicates that the three groups were not equivalent (p < .05) with respect to intelligence scores. Therefore, analysis of covariance method which partialled out pupils' intelligence scores on posttest results, was applied to analyze this difference.

TABLE 5

SUMMARY OF F VALUES FOR VARIABLES GROUPED ACCORDING TO SEX

	Male (n = 63)		Female (n = 53)		F Ratio
	$\overline{X}$	sd	$\overline{X}$	sd	
Age	84. 61	4. 63	84. 75	4. 62	0. 03*
IQ	104. 86	9. 92	106. 76	11. 86	0. 84*
SES	2. 38	0. 79	2. 58	0. 75	2. 02*
Pretest	10. 76	2. 22	10. 58	2. 72	0. 15*
Posttest	11. 59	2. 76	11. 70	2. 50	0. 05*

*Not significant at . 05 level of confidence.

At the . 001 level, an F ratio of 7. 54 (an approximate interpolated value) is needed for significance. The data cited in Table 6 shows the F ratio to be 10. 23. Therefore, a significant difference still existed among the groups for posttest scores, with the effects of intelligence scores partialled out. Again, as in the analysis of the data presented in Table 2, it was not possible to determine which group was significantly different from another without application of the Scheffe method.

The Scheffe method was applied and with pupils' intelligence scores partialled out, the findings from the posttest scores were:

1. Group A was significantly greater than Group B,

2. Group A was not significantly different from Group C,

3. Group C was significantly greater than Group B.

These findings indicate that pupils instructed with SRA materials for 2 semesters, and pupils instructed with Follett materials for 1 semester and SRA materials the other semester, achieved higher scores on the achievement test than pupils who were instructed with Follett materials all year.

TABLE 6

ANALYSIS OF COVARIANCE FINDINGS ON POST-TEST ACHIEVEMENT SCORES GROUPED ACCORDING TO INSTRUCTIONAL MATERIALS WITH INTELLIGENCE SCORES PARTIALLED OUT

Source of Variation	df	Corrected Sums of Squares	Corrected Mean Square	F Ratio
Instructional Materials	2	114. 33	57. 16	10. 23*
Error	112	625. 70	5. 59	

* Significant at the . 001 level of confidence.

Rival Hypothesis II. The differences noted were due to the fact that groups did not start at an equal point (as demonstrated in the pretest analysis), not due to the instructional program employed. Table 2 indicates that the three groups were not equivalent (p < . 01) with respect to pretest scores. Therefore, analysis of covariance method, which partialled out pupils' pretest scores on posttest results, was applied to analyze this significant difference.

An F ratio of 4. 89 (an approximate interpolated value) is needed for significance at the . 01 level. The data cited in Table 7 show the F ratio to be 7. 17. Therefore, significant differences still existed among the groups for posttest scores with the effects of pretest scores partialled out. Again, as in the analysis of the data presented in Table 2, it was not possible to determine which group was significantly different from another, without application of the Scheffe method. The Scheffe method was applied, and the findings from posttest scores, with pretest scores partialled out, were:

1. Group A scores were significantly greater than Group B scores;

2. Group A scores were not significantly different from Group C scores;

3. Group C scores were significantly greater than Group B scores.

These findings indicate that pupils instructed with SRA materials for 2 semesters, and pupils instructed with Follett materials for 1 semester and SRA materials the other semester, achieved higher scores on the achievement test than pupils who were instructed with Follett materials all year. These indications were also shown by the findings presented in Table 6.

Rival Hypothesis III. The differences noted were due to the differences which existed in the teacher's understanding of the substantive material taught (economics), and not due to the instructional material employed.

Although the data of Table 2 do not contain any tests relating to teachers' understanding of economics

TABLE 7

ANALYSIS OF COVARIANCE FINDINGS ON POST-TEST ACHIEVEMENT SCORES GROUPED ACCORDING TO INSTRUCTIONAL MATERIALS WITH PRETEST SCORES PARTIALLED OUT

Source of Variation	df	Corrected Sum of Squares	Corrected Mean Square	F Ratio
Instructional Materials	2	82.41	41.21	7.17*
Error	112	643.33	5.74	

*Significant at the .01 level of confidence.

as a variable, which could confound the results of the study, it seems that observation of the findings might pose the hypothesis that teacher understanding of economics would have an effect. As mentioned earlier, each pupil was assigned a score indicating his own teacher's level of understanding of economics. Teachers' level of understanding of economics was measured by the Test of Economic Understanding (TEU).

The data showed an F ratio of 6.69. For the F ratio to be significant at the .01 level, an F ratio of 4.89 (an approximate interpolated value) was needed. Therefore, the data showed that differences among the groups on posttest scores were still significant with the teachers' understanding of economics partialled out. However, as in the cases of the intelligence scores and the pretest scores, it was not possible to determine which group was significantly different from another without application of the Scheffe method. The Scheffe method was applied and the findings from posttest scores, with teachers' TEU scores partialled out, indicated:

1. Group A was significantly superior to Group B;

2. Group A was not significantly superior to Group C;

3. Group B was not significantly superior to Group C.

These findings indicate that pupils instructed with SRA materials for 2 semesters achieved significantly greater scores than pupils instructed with Follet materials all year. However, in this instance (TEU scores partialled out of posttest scores), pupils who were instructed with Follett materials for 1 semester and SRA materials the other semester did not achieve significantly higher scores than pupils instructed with Follett materials all year.

SUMMARY

For pupils' scores on achievement posttest, based on instructional materials, data resulting from this study indicate:

1. First-grade pupils of West Springfield, Massachusetts, who were instructed with SRA materials all year, achieved consistently higher scores on the measuring instrument than pupils instructed all year with Follett materials.

2. In two out of three instances (when intelligence scores and pretest scores were partialled out of posttest results), pupils who were taught with both Follett and SRA materials achieved higher scores on the measuring instrument than pupils instructed all year with Follett materials.

3. In one case out of three (when TEU scores of teachers were partialled out of posttest results), pupils who were instructed all year with Follett materials equaled achievement of pupils taught with Follett materials 1 semester and SRA materials the other semester.

For pupils' scores based on socioeconomic status, data resulting from this study indicate that no significant differences existed within and among the groups in posttest achievement when students were grouped according to social position.

For pupils' scores based on sex, data resulting from this study indicate that no significant differences existed among the groups in posttest achievement scores based on sex.

CONCLUSIONS

Based upon the data collected and analyzed in this study, the following conclusions were drawn:

1. Pupils instructed with SRA materials achieved consistently higher scores on a test of achievement in economics than did pupils who were instructed with Follett materials.

2. The socioeconomic status of the pupils did not effect performance on the achievement test. This result does not support the results of another study involving SRA materials which found that pupils of lower socioeconomic status performed at a lower level than pupils of middle socioeconomic levels.

3. A combination of instructional materials — 1 semester with SRA materials and the other with Follett materials-seems to result in a level of

TABLE 8

ANALYSIS OF COVARIANCE FINDINGS ON POST-TEST ACHIEVEMENT SCORES GROUPED ACCORDING TO INSTRUCTIONAL MATERIALS WITH TEU SCORES PARTIALLED OUT

Source of Variation	df	Corrected Sums of Squares	Corrected Mean Square	F Ratio
Instructional Materials	2	81.55	40.78	6.69*
Error	112	682.90	6.10	

* Significant at the .01 level of confidence

achievement in economics equal to that of pupils instructed with SRA materials all year.

4. The evidence provided by this study indicates that pupils can learn economics through an interdisciplinary approach.

IMPLICATIONS

This study did not attempt to assess the SRA or Follett materials "in toto." This study was concerned only with comparing the effectiveness of the materials on achievement in economics. However, these materials are concerned with much more than economics —geography, anthropology, political science, sociology, and history are interrelated parts of each set. Thus, a need exists to investigate how these interrelated parts work together —as a whole —to accomplish the purposes for which the materials exist.

Systems analysis asks the educator to see his activity as a whole —not only the instructional materials but also the child, the curriculum, the media, the teacher, and the management network which puts these and other resources together into a functional system. Educators might then acquire needed measurements on expenditure of energy and resources. Therefore, a new approach to the materials problem might be to think assiduously in terms of the way materials relate to the entire educational process.

Recommendations for Further Research. Recommendations based on the data and observations of this study are as follows:

(1) Using these first-grade social studies instructional materials, a full year study should be made to determine the effect of greater time duration on achievement in economics.

(2) A study should be made involving the Follett and SRA materials with groups selected by a random sampling method. As noted in the study, even with intelligence scores statistically equated, the pupils instructed with SRA for 1 semester and Follett the other semester, achieved scores on the achievement test significantly greater than pupils who were instructed with Follett materials all year. More data should be obtained on this variable, because the pupils instructed with Follett materials possessed significantly higher intelligence scores. In addition, the variable's "suppressing" or "moderating" effect on the posttest scores needs more investigation.

(3) A study should be conducted to assess more precisely the effect of teacher training on the teaching of these materials. "Teacher proof" instructional materials need much more empirical examination to warrant that label. Therefore, further research on the SRA and Follett materials should involve both teachers who have experienced in-service training in teaching economics, and teachers who have experienced no training at all in teaching economics.

(4) Both the instructional materials investigated in this study pursue the learning of economics through an interdisciplinary approach. A study which compared the achievement of first-grade pupils who were taught economics as an independent discipline, to pupils who were taught economics through an interdisciplinary approach, could provide evidence of the effectiveness of the two methods.

(5) As stated elsewhere in the study, both the SRA and Follett materials are concerned with attitudes and values. There has been virtually no research on the effectiveness of these materials in the areas of attitudes and values. It would be of considerable interest to investigage whether or not affective changes take place with pupils who have been taught with these materials.

(6) There is a need to assess these materials with urban pupils, because most of the studies which have measured first-grade pupil achievement in economics have focused on white suburban children.

(7) In order to confirm or reject the findings of this exploratory study, parts of it should be replicated in other sections of the country with new variables introduced. There is great need for more empirical evidence on other materials, other methods, and other objectives. The major problems of educational research are so big and so complex that breakdowns into minor problems might yield findings which are significant for solving the grand problem.

REFERENCES

1. Hollingshead, August B., Two Factor Index of Social Position, A. B. Hollingshead, New Haven, Conneticutt, 1958.

2. Jefferds, William J., A Comparison of Two Methods of Teaching Economics in Grade One, University of California, Berkeley, 1966.

3. Popham, W. James, Educational Statistics: Uses and Interpretations, Harper and Row, New York, 1967.

4. Spears, Sol, Children's Concept Learning in Economics Under Three Experimental Curricula, University of California, Los Angeles, 1967.

RESTRUCTURING HYPOTHESIS FOR

A PRESCRIBED SYMBOL CORRELATION

IN ALPHA-NUMERIC RECOGNITION

CHARLES T. ST. CLAIR[1]   KENNETH G. LEIB[2]   BENJAMIN J. PERNICK[2]

ABSTRACT

Alpha-numeric fonts have hitherto been designed with an aesthetic end in view. Acceptability for student reading material is achieved through statistically derived evaluations. Our tests show that optical spatial frequency spectra and cross correlation of font symbols can be quantitatively measured in the laboratory and that they provide a sound basis for differentiation among symbols. This enables the reader to note subtle differences between type fonts or within a given font. The method also allows the optical spectrum to be altered, and a modified or new letter to be constructed or restructured with improved character recognition for reading.

THE TEACHING of reading continues to receive major attention in educational research and development. This is due in part to the fact that ability to read well is basic to academic success and appears to be a controlling factor as to the vocation a student will follow and the type of life he will lead. Equally responsible for this interest in reading is the frustration educators feel at the puzzling lack of success in learning to read experienced by a large number of pupils, particularly boys. No one method and no single set of materials for the teaching of reading have thus far proved equally effective for all pupils. Thus research continues to focus on all aspects of the problem, with notable emphasis upon the reading process itself.

The primary goal in reading may perhaps be defined as the instantaneous transfer of thought from the printed page to the reader's mind through a visual and neurological process. It has been noted that the recognition of letter symbols by the eye is a function that frequently causes difficulty for beginning readers, especially with letters that represent reversals or inversions of characters of similar shape (p and q, m and w, d and b, etc.). The authors propose the hypothesis that a quantitative measure of letter symbols is derivable. The following questions immediately arise: Can a minimum quantitative standard of difference be established for letter symbols? Can type design be developed or revised for letters showing a high degree of similarity to make it easier for the eye to differentiate among them? Would an alphabet with greater differences in recognition among letter symbols be a help to pupils learning to read and prevent reading blocks that often afflict otherwise intelligent pupils?

The ability to quantify similarities and differences in recognition of letter symbols may be a useful tool in setting up experiments to answer these and similar questions in the reading field.

The brief review by Gray (2:1108-1111) considers many aspects of font and type size as functions of preference, prevalence, or statistical acceptance. It is the statistical aspect of the problem that the authors seek to reduce in the usually lengthy process required for a statistical derivation of meaningful letters and letter combinations for early-learning reading matter. Our preliminary study shows that the members of at least one alphabet (the IBM typewriter uppercase font) have inter-symbol uniqueness and measurable correlation.

METHOD

The Fourier transform properties of a coherent optical system are used to perform the measurements. The spatial frequency spectrum of a single letter symbol can be generated with an optical correlator (4:79-90). The letter symbol to be analyzed is contained in the system as a white-on-black film transparency. Monochromatic laser light is transmitted through clear portions of the film and collected with a transform lens. The spatial frequency spectrum of the letter symbol is realized as a light distribution pattern in the back focal plane of the lens. The zero frequency is on axis. Other frequencies in the two dimensional space are measured radially from the axis. A light detector place anywhere in

FIGURE 1

SCHEMATIC DIAGRAM OF OPTICAL SYSTEM FOR MEASURING CORRELATION OF TWO LETTER
SYMBOLS

TRANSPARENCY
OF FIRST LETTER

TRANSPARENCY
OF SECOND LETTER
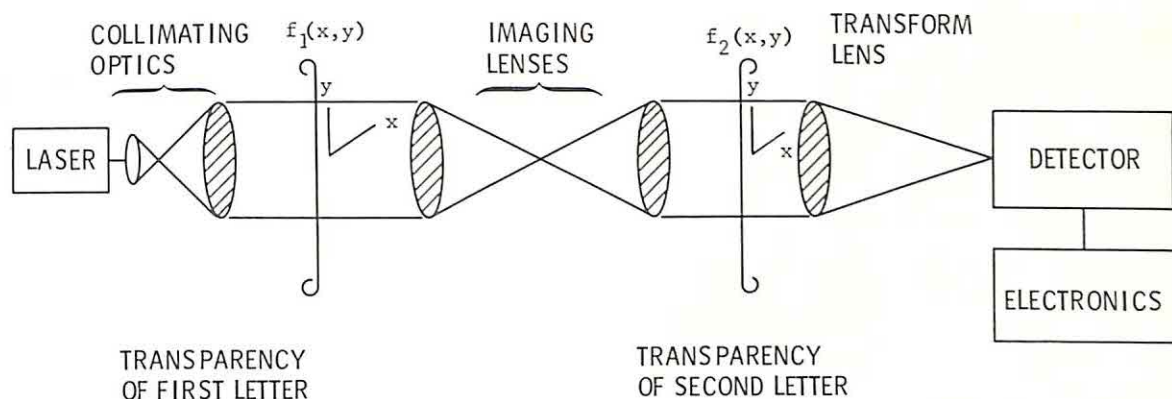
this plane measures the two dimensional light pattern so obtained.

A simplified arrangement of the optical components used to measure the correlation between two letter symbols is shown in Figure 1. Both letter symbols are presented as film transparencies. An image of the first symbol (B) is projected onto the second (W). Laser light that has been transmitted by both film strips is collected by the transform lens. The 2-dimensional light distribution pattern observed in the back focal plane of this lens now represents the spatial frequency spectrum of the superposed letters.

Let the real functions $f_1(x, y)$ and $f_2(x, y)$ represent the black and white density variations on the first and second film transparency, respectively. The light intensity I measured at points in the back focal plane is given by

$$I \approx \left| \iint f_1(x,y) f_2(x + \xi, y + \eta) \exp(j\nu x) \exp(j\lambda y)\, dx\, dy \right|^2 \tag{1}$$

where ξ and η refer to the x- and y-displacements of the first transparency with respect to the second. The parameters ν and λ are called spatial frequency variables and are proportional to the position coordinates in the back focal plane. The light intensity at the focal point, i.e., $\nu = 0 = \lambda$, becomes

$$I_o \approx \left[\iint f_1(x,y) f_2(x + \xi, y + \eta)\, dx\, dy \right]^2 \tag{2}$$

The measured intensity I_o is related to the square of the correlation of the letter symbols when both symbols are oriented at the same angle (4:89).

Restructuring either or both letter symbols results in changes in the correlation values between them. Altering the color or wavelength of the illuminating light also conveniently changes the frequency scale of the light intensity distribution. This is equivalent to increasing (or decreasing), without distortion, the size of the recorded letter symbol.

Optical correlation devices utilizing conventional (white) light sources measure only the correlation given by equation 1. A laser illuminator also makes possible measurement of the mathematical correlation and spatial frequency content of the letter symbols. This is a consequence of the color purity and coherence properties characteristic of the laser.

RESULTS

The letters B and W from an IBM typewriter font were chosen to illustrate the method. Typed symbols were photographed as negative transparencies; Figures 2 and 3 show the spatial frequency spectra obtained. Certain features observed in the spectrum are directly attributed to details of the letter symbol. For example, the main vertical bar of the letter B generates the horizontal pattern surrounding a bright central spot in the spectrum. Horizontal portions of this letter are responsible for the vertical frequency structure. Similarly, the horizontally-oriented serifs of the W yield the vertical structure in the spectrum. Slanted lines in this letter symbol generate the intricate patterns located on both sides of the central spot. The absence of a horizontal pattern in the W-spectrum is to be expected since this letter symbol has no vertical bar in its composition. Circular characteristics in a letter symbol yield a system of concentric rings in the spectrum. The degree of circularity is dependent upon the extent to which circular components are present in the letter. This can be seen in the spectrum of the letter B shown in Figure 2. The detailed composition of the spatial frequency spectrum depends on the length, thickness, and orientation of each bar, the number of parallel elements in the letter symbol, and the spacing between elements. The curvature of the letter and the number of circular parts also contribute to the spectrum.

Figure 4 shows the spatial frequency spectrum obtained from the superposition of the letters B and W. The intensity of the bright center spot is a measure of the correlation between the letters. Note that the spectral patterns surrounding the center spot

FIGURE 2

OPTICAL FREQUENCY SPECTRUM OF THE
ROMAN LETTER B
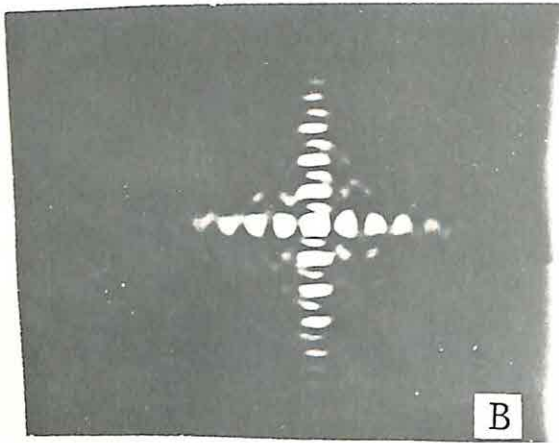
FIGURE 3

OPTICAL FREQUENCY SPECTRUM OF THE
ROMAN LETTER W
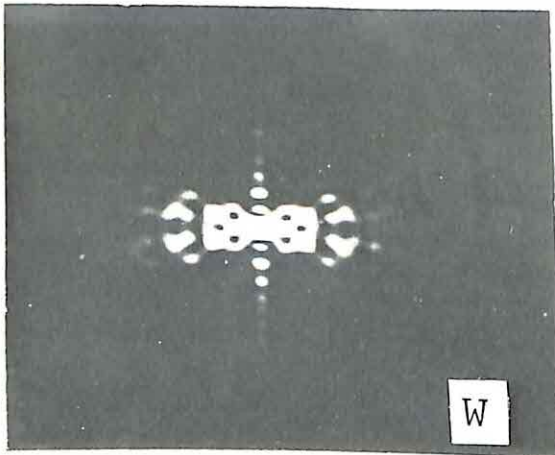
are more complicated for this B-W combination than
for the individual letter symbols alone.

DISCUSSION

Criteria for the design of an alphabet structure
should include the visual response of the eye. Mea-
surements of the relative response of the human vi-
sual system have been reported (3). Figure 5 shows
their frequency response curve for the eye as a func-
tion of spatial frequency on the retina. The relative
response factor is defined as the ratio of the ampli-
tude of the luminance variation to the mean value of
luminance.

The visual response curve exhibits a relatively
broad maximum at a retinal spatial frequency slight-
ly above 10 cycles/mm. The response has fallen by
two orders of magnitude at a retinal spatial frequen-
cy of about 200 cycles/mm. The decrease in re-
sponse at frequencies below 10 cycles/mm. is less
pronounced. Although these results are subject to

FIGURE 4

OPTICAL FREQUENCY SPECTRUM OF THE
CORRELATIVE COMBINATION OF ROMAN
LETTERS B AND W
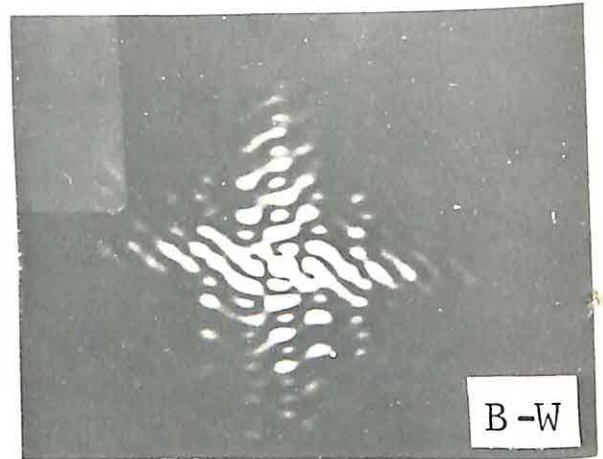
variations and uncertainties, they can be used as a
guide in a design program.

A program to revise or restructure an alphabet
based on prescribed correlative relations subject to
practical constraints should overcome many of the
shortcomings of the present alphabetic structure.
The spatial frequency content in the spectrum of a
given letter symbol should cluster about a value for
which the visual response is a maximum, i.e., about
10 cycles/mm. Experimentally, one attempts to
maximize the light falling in a circular ring sur-
rounding the zero frequency component of the ob-
served spatial frequency spectrum. The center ra-
dius of this ring corresponds to the peak spatial
frequency value of 10 cycles/mm. A preassigned
spatial frequency band determined by the width of
the circular ring would be a laboratory-controlled
parameter. Individual letter symbols should be
made distinct so that the correlation of any two re-
structured symbols would be low. A quantitative
measure of correlation is readily achieved by mea-
suring the light intensity at zero frequency of two
superposed letters, as described above.

Allowable levels of correlation would be a system
parameter. The alphabet should not, of course, be

FIGURE 5

SPATIAL FREQUENCY RESPONSE OF EYE IN
AVERAGE SAMPLE ACCORDING TO LOWRY AND
DE PALMA (1961)
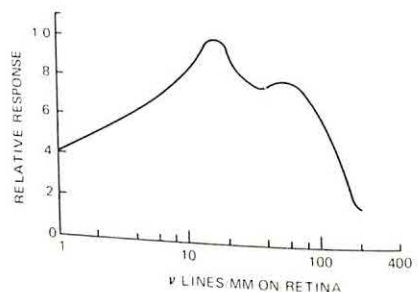
modified to such an extent that it no longer resembles existing styles, since this would be impractical and self-defeating. In view of the wide variety of existing fonts (1), a practical approach would be to select letter symbols and styles from those sources that satisfy the above constraints. One could, for example, analyze for spatial-frequency the Hoffman and other current alphabets designed for beginning readers to measure quantitatively how well they meet the guidelines stated.

FOOTNOTES

1. Superintendent of Schools, UFSD Number 23, Wantagh, New York 11793.

2. Research Department, Grumman Aerospace Corporation, Bethpage, New York 11714.

REFERENCES

1. Berry, W. T.; Johnson, A. F.; Jaspert, W. F.; The Encyclopedia of Type Faces, Pitman Publishing Company, New York, 1962.

2. Gray, W. S., in Harris, E. W. (ed.), Encyclopedia of Educational Research, Macmillan, New York, 1960.

3. Lowry, W. M.; DePalma, J. J., "Sine Wave Response of the Visual System: I. The Mach Phenomenon," Journal of the Optical Society of America, 51:740-746, 1961.

4. Stroke, G. W., An Introduction to Coherent Optics and Holography, Academic Press, New York, 1966.

BOOK REVIEWS   *Continued from page 48*

for the development of a mature, rational morality. "Psychological Standpoint" focuses on six underlying assumptions about children and learning. "Teachers and Children" provides a description of a day-long observation of one first school classroom in an attempt to provide situational settings for the theoretical aspects of the book.

Weaving throughout the book is a call for the availability of a variety of materials which can be used in a variety of environmental settings and which will allow a variety of children to select their own pace and method of learning and discovering.

But knowing Piaget does not justify the authors' existence. Their forte should be new insights into classroom application. This is where, as often happens, Piaget is used merely as a backdrop for someone else's ideas. A few examples should suffice.

To say, "teachers must rely less on apparatus and methods and more on the child himself—the knowledge, skills, and enthusiasm with which he arrives at school—in order to help him define and organize his responses" sounds great in a one-to-one or one-to-small-group ratio. The fact is that this knowledge is neither available about each individual child nor are teachers trained to be able to interpret it. Are we, therefore, left with the authors' "normal" description of stages of development which are not applicable to each individual child? In fact, is it not the provision of a wide variety and range of apparatus which lets children show us what their needs, abilities, and interests are?

The discussion of literature provides an interesting paradox between the personal impression and reflections the authors' emphasize and their didacticism. To say, "Wanda Gag's Millions of Cats is particularly absorbing to those children who have just developed the notion of one-to-one correspondence," puts this children's classic in a totally different context. That the "Ugly Duckling" is appreciated and enjoyed by children who have developed "conservation and reversability" is ridiculous when, in fact, this story may be a better example of having provided some children an opportunity to experience and begin to understand these two operations for the first time. The statement that "The House That Jack Built" does not demand "conservation of thought in the listeners" because of its repetitive nature raises some question as to whether the authors' fully understand the meaning of Piaget's terms.

Self concept is not dealt with systematically and this important idea continues to be left to chance even after a lengthy discussion.

Finally, the distinction is never made between children's incidental interests and the right and duty of the school to offer competing stimuli so that some directed development will take place. The emerging curriculum is only as good as the motivation behind it. Where that is shallow, other priorities and techniques for involving children in these priorities need to be brought into play. American schools need to look at the best of the British Infant Schools, but a wholesale transplant will only be effective when the purposes of education and the most effective ways of implementing them are in agreement on both sides of the water.

Oh, Piaget, if only you had developed a theory of teaching, as well as a theory of learning, we would not have to sort through these attempts to justify personal points of view.

Harlan S. Hansen, Reviewer
University of Minnesota

CONFIGURATION AS A CUE IN THE WORD

RECOGNITION OF BEGINNING READERS[1]

HENRY G. TIMKO

University of Victoria

ABSTRACT

Forty first-grade children responded to a delayed recognition matching—to—sample task involving 3—letter nonsense words. A 2X2X3 counterbalanced design involving sex, word shape (same and different), and shared letters (first, last, and none) was employed. Response words with the first letter or the last letter identical to that in the sample were chosen more often ($p < .01$) than responses with no elements identical to those in the sample. Words with the same first letters were confused more often ($p < .05$) than words which shared terminal letters. There was no significant main effect for shape, nor were there differences due to shape at any of the identical letter dimension levels.

ONE of the problems of continuing interest to researchers in the area of beginning reading has been the isolation of cues which are used in visually discriminating one word from another. Two cues which have been considered important are shared letters and similar word shape or configuration. However, research findings on the relative importance of these cues have been equivocal. While investigations by Levin and Watson (3), Levin, Watson, and Feldman (4); and Marchbanks and Levin (5) suggest that the first and last letters of words are the most critical cues in word recognition, other studies (1, 2, 7), emphasize the importance of word shape.

One of the most frequently cited recent studies on this topic has been that of Marchbanks and Levin (5). Kindergarten and first-grade children were given a delayed recognition matching—to—sample task containing trigram and quin-gram discriminations. On each presentation, the child was required to choose one of four response alternatives. Each response contained only one cue which was like that found in the sample. These cues were identical letters (first, middle, or last) or similar shape. The Ss most often confused words with identical first letters, and next most often confused those with the same last letters. Similar shape produced the fewest errors. Thus, it appears that when viewing these cues in isolation, identical first and identical last letters produce more confusion errors between words than does shared shape.

In Davidson's (1) investigation, which suggested that geometric shape is the most used cue, the experimental words included more than one cue at a time. For example, the author stated, "When party was confused with pretty the cause was laid to similar configuration, but not so when party and play were confused" (1:221). Were the Ss in that study confusing party and pretty because of similar shape, because of identical first and last two letters, or because of these cues in combination?

In an attempt to provide a methodological bridge between the Davidson (1) and Marchbanks and Levin (5) studies, the present investigation included not only a comparison of the relative saliency of identical letters and geometric shape, but also the main effect of these variables and their interactions on word recognition.

METHOD

Subjects

The Ss were forty randomly selected first—grade children from School District 61, Victoria British Columbia. There were an equal number of boys and girls in the sample. Confounding produced by formal reading instruction was minimized by conducting the study during the first few weeks of instruction.

Materials

The trigrams used in this investigation were

English letter nonsense words. The 1/4-inch letters were produced by a Primary typewriter. Long letters such as b, f, y, and p were extended to 1/2-inch to develop configurational differences in the words. The sample stimuli and response alternatives were typed on booklet—bound 5x8-inch cards.

A completely counterbalanced design employing the dimensions, identical letters (initial, terminal, and none), and shape (same and different) was used. Each response card contained three of the possible six response conditions. For example, if the sample was xrg, the response choices may be mug (identical terminal letter, same shape), ewj (no identical letters, same shape) and xds (identical initial letter, different shape). Each response condition occurred an equal number of times in combination with all others and each occurred twenty times throughout the forty trials. The various response conditions were randomly positioned on the response cards.

Procedure

The discrimination task included forty matching—to—sample delayed recognition trials. Each trial involved exposing a sample card on which there was a centered trigram. After looking at the sample stimulus for 5 seconds, the card was turned to reveal vertically positioned response alternatives. The S was instructed to put his finger on the word that looked to him most like the sample. During the first five trials, following each response, the experimenter provided verbal encouragement such as: " Right, Good, Fine. " No verbal feedback was given thereafter. When the S pointed to more than one response term in a given trial, only the first choice was recorded.

RESULTS

The number of choices given to each response condition is presented in Table 1. A 2X2X3 repeated measures design (6) with sex, shape, and identical letters as factors was conducted. There was no significant variation attributable to sex or word shape, nor were there any significant interaction effects. A significant main effect ($p < .001$) was found on the identical letter factor.

The Newman-Keuls comparison of means indicated that response alternatives with first letters identical to those in the samples were chosen more often ($p < .05$) than responses with identical last letters. The first letter and last letter conditions were each picked significantly more often ($p < .01$) than responses containing no letters identical to those in the sample.

DISCUSSION

The results of this investigation largely support the findings of Marchbanks and Levin (5). As in their study, the first letter in the word seems to be utilized more often by beginning readers than any other cue. And again, this tendency appears to be consistent across sexes. In addition, this study suggests that word shape, when studied as a main effect or as a possible contributor to interaction differences, is a relatively weak cue in this type of recognition task.

TABLE 1

CONFUSION ERRORS ON EACH RESPONSE CONDITION

| | Identical Letters | | |
	Initial	Terminal	None
Same Shape			
Boys	156	153	89
Girls	163	152	94
TOTAL	319	305	183
Different Shape			
Boys	182	129	92
Girls	166	144	80
TOTAL	348	273	172

It appears that prior to formal reading instruction, children are attending more to the features of letters in words than they are to total word shape. This finding is of particular relevance to the practice of geometric and word—shape training in pre—reading programs. Since children exhibit a strong tendency to focus on the individual letters of words they are to learn, does training to attend to another cue such as word shape facilitate or inhibit their discrimination learning? Additional research on the transfer value of configuration training seems to be in order.

REFERENCES

1. Davidson, P., " An Experimental Study of Bright, Average, and Dull Children at the Four—Year Mental Level, " Genetic Psychology Monographs, 9:119-225, 1931.

2. Foote, W. E.; Havens, L. L., " Stimulus Frequency: Determinant of Perception or Response," Psychonomic Science, 2:153-154, 1965.

3. Levin, H.; Watson, J.," The Learning of Variable Grapheme-to-Phoneme Correspondences," Final Report, United States Office of Education, Basic Research Program on Reading, 1963.

4. Levin, H.; Watson, J. S.; Feldman, M., " Writing as Pretraining for Association Learning," Journal of Educational Psychology, 55:181-184, 1964.

5. Marchbanks, G.; Levin, H., " Cues by which Children Recognize Words," Journal of Educational Psychology, 56:57-61, 1965.

6. Winer, B. J., Statistical Principles in Experimental Design, McGraw-Hill, New York, 1962.

7. Woodworth, R. S., Experimental Psychology, Holt, Rinehart, and Winston, New York, 1938.

TEACHING OBJECTIVES, STYLE, AND EFFECT

WITH THE CASE METHOD IN ENGINEERING

KARL H. VESPER
University of Washington

JAMES L. ADAMS
Stanford University

ABSTRACT

This article seeks to trace relationships between teaching objectives, teaching style, and emphasis as perceived by students in an engineering course taught by four different instructors using the case method of instruction. A checklist was used for gathering data regarding objectives of the instructors and effects perceived by the students. Data from tape recordings were used to make inferences about teaching style.

Experimental data for this study were gathered in the first year graduate course, "Case Studies in Mechanical Engineering" offered at the University of California, Berkeley, in 1968. This course has been offered once a year since 1965 under the direction of Professor Robert Steidel. It is taught by four people, each conducting approximately 2 weeks of it. The same four people taught it from 1965-1968. The catalogue description of the 1968 course was:

This course will introduce the case history and the case study as a means of experiencing mechanical design. Four or five selected case studies will be reviewed and discussed. These will involve the thermal control of a spacecraft, the machine design of an oil well drilling mechanism, a bearing problem, the control of a high speed centrifuge, and the structural design of a solar panel. Each will be critically analyzed to study the design process and the engineering decisions that were involved.

This course is open to all graduate students in Mechanical Design, but it is required of all candidates for the Master of Engineering degree who expect to undertake an engineering case study in lieu of a design project or thesis in the Spring Quarter, 1968.

Five cases were used in the 1968 course, one by each of the first three instructors and two by the fourth. The first one used was a case problem and the rest case histories. The case problem seeks to put the student into the position of an engineer facing an un-

solved problem, while the case history describes both the problem and its outcome so that the student can view both in retrospect (for further description see reference 1). All cases described real situations taken from industrial settings.

No requirements were imposed upon the individual teachers as to what cases they used or how they used them except for the general time limitations of the course itself.

TEACHING OBJECTIVES

For gathering data regarding apparent intrinsic aims of the course a questionnaire in the form of a "Teaching Objectives Checklist" was used. This checklist (3) was designed to elicit ratings of relative perceived emphasis for a variety of possible teaching objectives. A copy of this checklist was given to each instructor at the beginning and again at the end of his section of the course and to each student at the end of each section. Instructors and students were requested to indicate on the checklist the relative emphasis on each objective.

Before-and-after estimates by each of the instructors are shown in Table 1. Teacher 2 did not completely fill out the checklists because of lack of sympathy with the testing. As expected, the checklists showed a difference between instructors and between before-and-after objectives estimates for each individual instructor. Simple statistical tests were used to show the relative significances of the differences between teaching objectives seen by each instructor, between before-and-after estimates by each instructor, and between teaching objectives estimates of the

four instructors. First of all, the variance between the before-and-after estimates was compared to the variance between the "after" emphasis estimates for each teacher. Applying an F-test to these variances indicated that for each teacher (excepting number 2, who did not complete the checklist) there were significantly higher (at the .05 level) variations in indicated emphasis between checklist items than between before-and-after estimates. If the before-and-after differences are considered as "noise" and the differences in emphasis are considered as "signal," the signal-to-noise ratio for the instructors is as follows: (the threshold of statistical significance is 1.84)

Instructor	Signal/Noise
1	3.78
2	—
3	1.93
4	4.82

Another analysis of variance test was performed on the "after" estimates of the various teachers in order to check for similarity of emphasis estimates between teachers. The F ratio (1.026) did not reach the threshold of significance (1.65). This test therefore did not show statistically significant agreement of estimated objectives between teachers. It is nevertheless of interest to note the items which drew the strongest indications of emphasis and non-emphasis. These were as follows:

Mean Rating	Checklist Item

Indicated Highest Emphasis

4.3	Spotting key facts amid less relevant data
4.0	In prescribing action to be more specific
4.0	Identifying and defining practical problems
4.0	Knowledge of what engineers do and how they work

Indicated Lowest Emphasis

1.5	Emphasis on reasoning quantitatively
2.0	Formulating idealized mathematical models
2.0	Manipulating and solving mathematical models
2.0	Aesthetic sensitivity

Although these results are not statistically significant (only three professors were polled) they show considerable agreement with statistically significant

results described elsewhere (3) which were obtained with a sample of thirty professors.

TEACHING STYLES

The four instructors in this course differed in homework assignments, conduct of class meetings, and type of questions asked. Homework assignments were roughly as follows:

Teacher	Assignment
1	Take up the engineer's job at the end of the case and carry it further.
2	Describe how you would follow or depart from the method of operation of the case engineer.
3	Assume the role of a project engineer who is not mentioned in the case and formulate plans.
4	Map the main decisions and events in the case in the form of a flow chart.

All teachers used discussion rather than lectures. However, from tape recordings it was found that classroom style varied measurably among the teachers. Table 2 shows pertinent quantities for each teacher during a sample class period (the second 15 minutes of discussion).

Table 3 classifies the statements of the teachers into seven categories, the first four of which ask the students to comment.

STUDENT RATINGS OF EMPHASIS

Table 3 shows the average relative rankings given to each objective by the students. Number 1 represents the objective most emphasized in the opinion of the students and number 30 represents the one least emphasized. A Q-test for differences between many pairs of means was used to test differences between higher and lower ranked objectives. The result was that the four highest ranked items for each teacher has statistically significantly (at the .051 level) higher mean ratings than the four lowest. Objectives ranked closer to each other could not be ordered with this confidence. Items thus significantly separated have been marked with an "H" for high and an "L" for low in Table 1.

RELATIONSHIPS BETWEEN OBJECTIVES, STYLE, AND PERCEIVED EMPHASIS

It is of interest to infer a few relationships between teaching objectives as rated by the instructors, teaching style as measured from tape recorder data, and emphasis as perceived by the students. It is not possible to draw rigorous conclusions because of the many variables (personality, organization) not covered by the testing techniques and because of the small sample of teachers. However, inferred relationships are of interest because they demonstrate the type of conclusions which can be drawn from this type of testing. Such relationships are useful not only to the teacher who wishes to evaluate his techniques or reorient his course, but also to those

TABLE 1

SUPERPOSITION OF STUDENTS' AND TEACHERS' RATINGS*

LOW ← scale (1 2 3 4 5) → HIGH (H = high, L = low)

Item	Teacher 1 (1–5)	Teacher 2 (1–5)	Teacher 3 (1–5)	Teacher 4 (1–5)
I. HABITS — Emphasis on increasing your tendency to				
1. Reason quantitatively (use numbers) whenever possible	A:2 B:3		A:4	B:1 A:1
2. Discriminate between fact and opinion	A:4 B:5 · H	H	B:3 A:4	B:1 A:2
3. Search for more alternative solutions	A:3 B:5 · H	H	B:2 A:3	A:2 B:4
4. In prescribing *action* be more specific	B:4 A:5		A:4	A:2 B:3
5. Pay meticulous attention to detail	B:3 A:4		B:3 A:4	B:1 A:1
6. Think more carefully before speaking	B:3 A:4		B:2 A:3	A:1 B:1
II. SKILLS — Emphasis on developing your ability of				
1. Using unfamiliar exercise tools or exercise methods	B:1		B:3 A:4	
2. Communicating (writing, speaking, or drawing)	B:3 A:4 · H	B:4 · H	B:2 A:3	B:1 A:2
3. Identifying and defining practical problems	A:2 B:3	B:3	B:4 A:5 · H	B:3 A:4 · H
4. Spotting key facts amid less relevant data	A:4 B:5	B	B:4 A:5 · H	B:3 A:4 · H
5. Foreseeing consequences of alternative actions	B:3 A:4		A:2 B:3	B:2 A:3 · H
6. Formulating idealized math models of real problems	B:1 A:1 · L	B:1 · L	B:3 A:4	B:1 · L
7. Manipulating and solving idealized math models	B:1 A:1 · L	B:1 · L	A:4 B:5 · L	B:1 · L
8. Viewing problems and complications with perspective	A:3 B:4		B:2 A:3	B:1 A:2 · L

Item				
9. Selling ideas or arguing more persuasively	B A			B A
III. KNOWLEDGE – Emphasis on imprinting on your memory				
1. Historical episodes as lessons	B A	B A	A B	B A
2. What engineers do and how they work (typical activities)	B A	B A	B A	B A (H)
3. Criteria for judging engineering procedures	A B	B	A B	B / A
4. Criteria for evaluating designs	B A	B	B A	B A
5. Mathematical, scientific, or engineering theory	B A	B	B A	B A
6. Technical facts used in engineering (besides theory)	B A		B A	B A
7. Formats or class procedures unique to the class	B A		B A	
IV. ATTITUDES AND FEELINGS – Lifting your level of				
1. Self confidence	B A		A B	B A
2. Perserverance in spite of setbacks	B A		A B	A B
3. Concern with questions unanswered for you yet	B A		B A	B A
4. Enthusiasm, motivation for course or engineering	B A			
5. Aesthetic sensitivity	B A		B A	A B
6. Interpersonal sensitivity (feel for people)	A B		A B	A B
7. Self knowledge (own abilities and limitations, etc.)	B A		B A	A / B
8. Tolerance for others ideas and errors	B A		B A	A B

Scale markers H (high) and L (low) appear across the rating columns.

* B signifies teachers' "Before" prediction; A signifies teachers' "After" prediction; H and L are student assessments from Table 4.

TABLE 2

STATISTICS ON SECOND FIFTEEN MINUTES OF CLASS DISCUSSIONS IN BERKELEY CASE COURSE*

	Teacher 1		Teacher 2		Teacher 3		Teacher 4	
Session	1	2	1	2	1	2	1	2
Number of Interchanges**	45	40	21	36	33	25	33	25
Time Teacher had Floor	44%	29%	37%	34%	72%	80%	44%	33%
Time Students had Floor	56%	71%	63%	66%	28%	20%	56%	67%
Teacher Average Time per Interchange (seconds)	8.3	5.4	14.2	7.5	18.2	26.9	11.4	10.8
Student Average Time per Interchange (seconds)	10.7	15.3	24.6	14.9	7.1	6.6	14.3	22.4
Total Words by Teacher	906	839	591	551	1508	2020	604	428

* Based on tape recordings of the discussions.
** Occasional brief interjections such as "yes," "OK," "I see" were not counted.

TABLE 3

TYPES OF COMMENTS BY TEACHERS IN CLASS DISCUSSIONS OF BERKELEY CASE COURSE

Type of Statement by Teacher*	Fraction of Teacher's Total Statements			
	Teacher 1	Teacher 2	Teacher 3	Teacher 4
1. Asked what engineer(s) in case should do next	16%	0%	6%	0%
2. Asked student to clarify or justify something student had said	37%	28%	0%	14%
3. Requested specific facts from case	7%	5%	7%	3%
4. Asked for student appraisal of what engineers in the case had done	9% (69%)	24% (57%)	9% (22%)	3% (20%)
5. Summarized or reiterated student comments	20%	19%	21%	16%
6. Gave directions	0%	10%	15%	30%
7. Stated own opinion or rhetorical description	11%	14%	42%	34%
TOTALS	100%	100%	100%	100%

* Classified subjectively by the experimenter based on tape recordings of the class discussions.

seeking to design new courses or curricula which include case studies.

During the planning of the course, Professor Steidel attempted to schedule the four teachers so as to begin with the most nondirective and increase the degree of directiveness as the term progressed. There are several indications that he succeeded. The most notable are the increasing ratio of directions given to total statements and the decreasing ratio of questions to total statements seen in Table 3 (69% questions for teacher 1, 57% questions for 2, 22% for 3, 20% for 4). Looking at the type of question asked, it can be seen that instructor 1 most often asked what the engineer(s) in the case should do next. This is consistent with his indicated emphasis on

TABLE 4

RANKS OF STUDENT PERCEIVED EMPHASIS ON OBJECTIVES OF EACH BERKELEY CASE COURSE TEACHER*

	Teacher 1	Teacher 2	Teacher 3	Teacher 4
I. HABITS - Emphasis on increasing your tendency to				
1. Reason quantitatively (use numbers) whenever possible	24	15	22	27
2. Discriminate between fact and opinion	17	11	17 18	18 21
3. Search for more alternative solutions	2	3	4	7
4. In prescribing <u>action</u> be more specific	1	4	20	16
5. Pay meticulous attention to detail	25	21 22	19	17
6. Think more carefully before speaking	8	25 26	17 18	22 23
II. SKILLS - Emphasis on developing your ability of				
1. Using unfamiliar exercise tools or exercise methods	23	25 26	13 14	18 21
2. Communicating (writing or speaking or drawing)	3 4	16 17	24	11 13
3. Identifying and defining practical problems	10	1	1	2
4. Spotting key facts amid less relevant data	21	9	3	1
5. Foreseeing consequences of alternative actions	5 6	20	5 6	3
6. Formulating idealized math models of real problems	29	27	25	29
7. Manipulating and solving idealized math models	30	30	26 27	30
8. Viewing problems and complications with perspective	11	12	10	9
9. Selling ideas or arguing more persuasively	3 4	24	30	25
III. KNOWLEDGE - Emphasis on imprinting on your memory				
1. Historical episodes as lessons	27	18 19	11 12	10 11
2. What engineers do and how they work (typical activities)	14	2	2	4
3. <u>Criteria</u> for judging engineering procedures	13	5 7	9	6
4. Criteria for evaluating designs	17	5 7	5 6	8
5. Mathematical, scientific, or engineering theory	28	28	15	28
6. Technical facts used in engineering (besides theory)	12	9	7	18 21
7. Formats or class procedures unique to the course	21	21 22	29	26

TABLE 4   (Continued from previous page)

	Teacher			
	1	2	3	4
IV. ATTITUDES AND FEELINGS - Lifting your level of				
1. Self confidence	5 6	13 14	26 27	15
2. Perseverance in spite of setbacks	21	18 19	21	11 13
3. Concern with questions unanswered for you yet	17	23	155	22 28
4. Enthusiasm, motivation for course or engineering	8	16 17	8	5
5. Aesthetic sensitivity	26	29	23	24
6. Interpersonal sensitivity (feel for people)	17	5 7	28	18 21
7. Self knowledge (own abilities and limitations, etc.)	17	9	13 14	10 11
8. Tolerance for other's ideas and errors	8	13 14	155	11 13

* Two rank numbers appearing for a given item indicate ranges of ties in rank

"increasing student's tendency to be more specific in recommending action" on the checklist. Instructor 3 was the only one who did not ask students to clarify or justify something they had said. From looking at Table 3, it can also be seen that instructor 3 most often "stated own opinion or rhetorical description." Table 2 indicates that instructor 3 held the floor much more than the other instructors (about 75% of the time, as opposed to approximately 35% of the time). These characteristics of style are consistent with unsolicited student complaints such as "talks too much," "tends to be too opinionated," "answers his own questions without exploring class."

Instructor 1 used a case problem in which each chapter ended in a clear problem situation involving a particular engineer. Instructors 2, 3, and 4 used case histories. These choices are consistent with the high emphasis upon "historical episodes as lessons" checked by instructors 2, 3, and 4. They are also consistent with many of the emphasis categories checked by the students. Instructor 1, for instance, was the only one who was not seen as emphasizing "identifying and defining practical problems." This was probably because the case identified and defined them specifically. Similarly, students probably did not rate teacher 1 high on "what engineers do and how they work" because his case was more problem oriented than description oriented.

Teacher 3 was the only one who was not rated lowest at "manipulating and solving idealized math models." He also was rated higher at conveying "mathematical, scientific, or engineering theory." This was possibly because his case was the most theoretically oriented of any of the cases. The previously mentioned tendency of teacher 3 to do more of the talking himself might explain the low emphasis on

teaching students to "argue more persuasively or to sell ideas" as perceived by his students. Similarly, the high rating of teacher 1 perceived here is consistent with his high rating of "communicating" and "selling ideas" and with his tendency to ask students more frequently to explain and justify their statements.

The concurrence of teacher objective emphasis and student perceived emphasis varied from teacher to teacher. Considering the six significant student ratings (marked in Table 1), it can be seen that teacher 4 was most accurate in agreement between his estimates and the perceived student emphasis. Teacher 4 also was best able to discriminate between items on the checklist according to the previously discussed "signal to noise" test. Teacher 1 agreed on only three out of six and teacher 3 on none out of six. (He put one at the top of the scale end—"selling ideas"—which students put at the opposite end.) Teacher 2, although he did not participate in most of the testing, agreed with students on the only item he rated high—"what engineers do." Since only six items drew significantly distinguishable ratings from the students, many of the teachers' estimates were uncheckable. Several items checked by the teachers at extremes of the scale can be considered as "misses", since the checklist was not sensitive enough to detect significant student opinion. Teacher 1 had six of these misses, teacher 2, five, teacher 3, four, and teacher 4, nine. Had the checklist been truncated by removing the highest contrast items and reapplied, enough further significant contrasts might have emerged to turn some of these "misses" into "hits." However, no time was available to do this. In the opinion of the investigators, a more significant line of further investigation would have been an attempt by the instructors to use the checklist to "steer" toward different rating patterns

as an indication of their degree of control over the teaching process with cases.

REFERENCES

1. "The Preparation and Uses of Cases in Engineering Education—A Call for Proposals," Commission on Engineering Educa-tion, Washington, D. C. March 1, 1963.

2. Department of Mechanical Engineering, University of California, Berkeley, Course Description (Mimeograph), 1968.

3. Vesper, K. H.; Adams, J. L., "Evaluating Learning from the Case Method," Journal of Engineering Education, 60: (no. 2) 104, October 1969.

THE RELATIONSHIP OF ACHIEVEMENT

RESPONSIBILITY TO INSTRUCTIONAL

TREATMENTS

KINNARD WHITE
University of North Carolina, Chapel Hill

JAMES LEE HOWARD
North Carolina Advancement School, Winston-Salem

ABSTRACT

The hypothesis that students who are characterized as having external versus internal control would achieve differently under different classroom instructional treatments was tested. Thirty-two seventh-grade boys (sixteen externals and sixteen internals) were randomly assigned to two instructional treatments in science. The treatments varied according to the amount of control exercised by the teacher. The role-assumption treatment emphasized student-directed learning experiences, whereas the structured treatment had maximum teacher control. The treatments were applied for 16 weeks. Analysis of covariance confirmed the hypothesis. The significant interaction indicated that internals achieved at the same level under both treatments, whereas externals achieved more under the role-assumption treatment.

RECENT research designed to increase our knowledge of the relation of individual differences to the learning of school children has demonstrated the importance of the variable internal-external control of reinforcement (8, 11). Rotter's (12) social learning theory postulates that the occurrence of a behavior in a particular circumstance is a function of the individual's expectancy that the behavior will result in reinforcement. The control construct in Rotter's theory is considered to be a generalized expectancy, operating across a large number of situations, relating to whether the individual accepts responsibility for what happens to him. Persons characterized as having internal control are considered to perceive events as being a consequence of their own actions and therefore in some sense under their personal control. Persons characterized as having external control are considered to perceive events as being unrelated to their own behaviors and, consequently, beyond their personal control.

Crandall, Katkovsky, and Crandall (2) have developed a scale (Intellectual Achievement Responsibility-IAR) to measure children's beliefs in responsibility for reinforcement. This scale is primarily devoted to school-related situations. Rather than focus on impersonal social forces, the IAR scale has limited the sources of external control to persons such as teachers, parents, and age mates, who are most likely to come into contact with the school-aged child.

Recent research on learning in laboratory settings using internal versus external control of reinforcement as a variable has shown that the internal person learns more rapidly, is less variable in his learning, generalizes his learning more, and retains more (8). Using the IAR scale, Crandall, Katkovsky, and Preston (3) found that achievement-related activities were highly related to control among boys. These relationships did not hold for girls. Specifically, boys who attributed responsibility for achievement to themselves (internals) spent more time in intellectual free-play, demonstrated greater intensity of striving in intellectual free-play, and scored higher on intelligence tests, reading achievement tests, and arithmetic achievement tests.

The evidence indicates that students characterized as externals do not perform as well as those characterized as internals when exposed to conventional school learning situations.

Jackson's (7) analysis of this situation has led him to suggest that the teaching-learning situations in the schools could be reconstructed so that students characterized as externals would be forced to view the teaching-learning environment not as one in which the student is performing in certain prescribed ways designed to elicit reinforcements from the teacher. If the external student is placed in a situation in which his task is to explore an area structured by himself,

then more positive results could be expected. Rogers (10) and Maslow (9) have also argued from motivational theory that the child must develop toward autonomy and away from external control.

Current trends in teaching and curriculum have emphasized both the individualization of instruction and the use of inquiry methods to provide for differential rates of learning (13). Actual attempts to implement such methods have met with varying degrees of success. Thelen (15) has suggested that one likely source of difficulty is the lack of knowledge of how teaching methods and student characteristics interact. Although theory has emphasized the nurturing of productive thinking through the design of learning experiences, application has lagged (6, 18). Furthermore, very rarely have researchers applied motivational theory to the design and conduct of instructional programs.

Bettelheim (1) has argued that the goals of enhancing psychological health and effective functioning are primary to the learning process. In addition, the skill of emotional management is dependent upon the development of autonomy and inner freedom within the individual. In developing his argument, Bettelheim emphasizes the following points:

First, the school must help the child overcome his lack of identity. This may best be accomplished by allowing the child to face conflicts rather than avoid them.

Second, teachers must develop the skill of self-direction in children by allowing them to experience life first-hand. This will not be accomplished by prepackaged explanations of what life (or school subject matter) is all about from the viewpoint of the teacher. Furthermore, the child should be encouraged to express his own point of view.

Third, the school must provide opportunities for the individual to manage his anxieties and to understand his own emotional reactions.

Fourth, the school should allow a child to encounter potentially dangerous external (failing) situations directly and without condemnation, rather than hiding the danger or purposefully reinforcing the failure externally. In this way, the child learns internally to identify the potential danger in a situation and is allowed to learn from his failures rather than to avoid or dismiss them.

The theoretical views on school curriculum and teaching expressed by Bettelheim, Jackson, Hullfish, Siegel, Smith, and Thelen echo the argument put forward by Cronbach (4). Cronbach's argument was that not only are treatments characterized by many dimensions, but so also are persons. Consequently, we must attempt to deal with treatments and persons simultaneously. In doing this, we must attempt to design treatments to fit groups of students with particular personality characteristics.

The research reported here was designed to test three hypotheses derived from the theory and empirical research data reviewed above. The hypothesis of primary interest was that the perceptions of male students with regard to internal versus external control of reinforcement in intellectual achievement situations would interact with instructional method in affecting achievement. Specifically, the following interaction hypothesis was tested: Achievement for externals would be superior under a method of instruction that stressed learner-directed activities as compared to a method of instruction that stressed teacher-directed activities: achievement for internals would not be affected by these differential instructional methods. In addition to the interaction hypothesis, which was of major interest, the following hypotheses were tested: (a) internals will achieve more than externals, and (b) students in the learner-directed group will achieve more than students in the teacher-directed group.

METHOD

Subjects

The Ss for this study were thirty-two boys enrolled in the fall 1968 term of the North Carolina Advancement School. All Ss had elected to take a general science course designed for seventh-grade students. The Advancement School is a residential school designed to conduct research on discrepant achievement. It is maintained by the state and selects students from public schools throughout the state. To be eligible, a boy must have average or above average ability and be achieving two or more grade levels below the school grade in which he is currently placed. Both of these criteria, ability and achievement, are determined by standardized tests. Most students are permitted to remain at the school for only one semester (16 weeks). All Ss were in the seventh grade and were enrolled in the Advancement School for the first time.

Design

A 2X2 factorial design with a pretest and posttest was used for the study. There were two levels of instructional method—role-assumption and structured class; reinforcement control consisted of two levels, internal and external.

Procedure

Assigned Variables. The Ss were categorized as being externals or internals on the basis of their scores on the IAR scale (2). This scale was routinely administered to all boys entering the Advancement School as a part of the testing program during the first 2 days in residency. The median score on the IAR scale for those who elected to take the science course was used to establish categories of internals and externals (median for the total group was 25). Those above the grand median were considered to be internals (median on IAR for this group was 28), and those below the grand median were considered to be externals (median on IAR for this group was 21).

Manipulated Variables. The treatment which stressed learner-directed activities was called a role-assumption treatment. The sixteen students assigned to this treatment were told that they were to assume the role of a scientist and could spend their time during the semester studying anything they desired as long as it was related to science. The role of the teacher was one of helping the student to secure information and materials for experiments or other ex-

plorations in which the student was interested. Students in this treatment performed two or more experiments during the semester. These experiments included such things as weather prediction, the effect of types of light upon plant growth, genetics, nutrition, and animal care.

The treatment which stressed teacher-directed activities was called the structured treatment. The sixteen students assigned to this treatment were given considerable direction as to the kinds of experiments likely to be of value in their learning. They were also given specific reading assignments for which they were held responsible. Except for grades, which were not given to students in either group, students in the structured group were, in every way possible, led to view the teacher as being in control of the learning situations and passing out reinforcements when they were justified.

The same two teachers taught both classes. Students were assigned to the treatments randomly, with eight students in each cell of the 4-cell design.

Criterion and Covariates. The criterion was achievement as measured by the Sequential Test of Educational Progress (STEP): Science (5).This test was selected as the criterion measure for three reasons. First and most importantly, it was designed to measure the skills in science that were the primary instructional objectives of the science course taught to both treatment groups for this research. The skills measured by this test are: identifying scientific problems, developing hypotheses, selecting valid procedures, interpreting data, critically evaluating claims, and quantitative and symbolic reasoning. Secondly, the content areas of this test are approximately equally divided into biological, earth (including astronomy and meteorology),and physical science. This, too, reflected the content range of the course. Thirdly, a standardized test was desired.

Two control variables were used, control being exercised by means of the analysis of covariance. Intelligence (14) and the pretest STEP-Science achievement were used as covariates.

RESULTS

The means and standard deviations on the covariates (IQ and pretest achievement) and the means, standard deviations, and adjusted means for the criterion (achievement) may be observed in Table 1.

The results of the analysis of covariance of the posttest achievement scores with IQ and pretest achievement as covariates may be observed in Table 2.

TABLE 2

ANALYSIS OF COVARIANCE FOR POSTTEST ACHIEVEMENT WITH PRETEST ACHIEVEMENT AND IQ AS COVARIATES

Source	df	ms	f
Treatment (T)	1	112.74	6.72*
Responsibility (R)	1	0.16	0.01
TXR	1	109.07	6.50*
Regression	2	73.03	4.35*
Error	26	16.78	

* $p < .02$

The hypothesis predicting an interaction between instructional method and internal-external type was upheld by the results of this analysis. The interaction was significant ($F = 6.20$, $df = 1, 26$, $p < .02$) and in the predicted direction. The type of interaction obtained is shown in Figure 1.

The level of achievement by students who were classified as internals was not affected by the method of instruction ($\overline{X}_{adj}$ role-assumption was 24.57; $\overline{X}_{adj}$ for structured was 24.89). However,achievement by students who were classified as externals was quite clearly affected by the method of instruction ($\overline{X}_{adj}$ for role-assumption was 28.62; $\overline{X}_{adj}$ for structured was 20.54). This was predicted by the hypothesis: internals did equally well in both instructional methods,presumably because they brought to the learning situation, regardless of how it was structured, the belief that they were responsible for their reinforcements. However, the achievement of externals, who did not bring to the learning situation the belief that they were responsible for their reinforcements,was differentially affected by the instructional method. The instructional method which forced the student to take command of the learning situation, in contrast to the instructional method which the student could clearly perceive as one in which the teacher was the source and controller of reinforcements,resulted in superior achievement for this type of student.

TABLE 1

MEANS AND STANDARD DEVIATIONS FOR COVARIATES AND MEANS, STANDARD DEVIATIONS, AND ADJUSTED MEANS FOR EACH CELL (N = 8) IN THE EXPERIMENTAL DESIGN

	INTERNALS						EXTERNALS					
	Role			Structure			Role			Structure		
Variable	s	$\overline{X}$	$\overline{X}_{adj}$	s	$\overline{X}$	$\overline{X}_{adj}$	s	$\overline{X}$	$\overline{X}_{adj}$	s	$\overline{X}$	$\overline{X}_{adj}$
IQ	10.30	96.12		9.20	102.88		16.31	104.00		10.61	96.62	
Ach (Pre)	4.75	19.62		3.33	19.25		5.68	22.38		4.31	20.50	
Ach (Post)	4.59	23.25	24.57	4.64	24.88	24.89	4.78	30.00	28.62	4.21	20.50	20.54

FIGURE 1

GRAPH OF REINFORCEMENT TYPE BY TREAT-
MENT INTERACTION (F = 6. 50, df = 1, 26, p < . 02)

_____ Role-Assumption
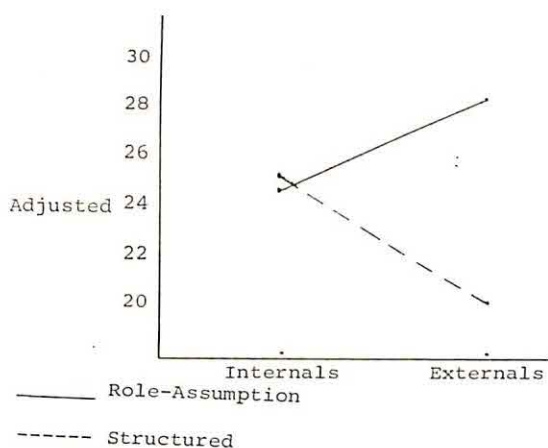
------ Structured

There was a main effect for the treatment factor
(F = 6. 72, df = 1, 26, p < . 02). Students assigned to the
role-assumption treatment group achieved much bet-
ter than those assigned to the structured treatment
group. The adjusted mean for posttest achievement
for those in the role-assumption treatment was 26. 60,
and 22. 72 for students in the structured treatment.
The hypothesized difference in achievement between
externals and internals was not supported (F = 0. 01,
df = 1, 26, p was nonsignificant).

DISCUSSION

The results clearly indicate that it is possible to
design treatments to match personality types of stu-
dents to improve achievement (4). Under the normal
school learning situation, the structured treatment,
external students did not perform well on the science
achievement test. However, when placed in a method
of instruction designed especially to fit their person-
ality type, their achievement was improved to a level
considerably above that of internal students. Such re-
sults should be encouraging to designers of instruc-
tional methods and should alert applied psychologists
to the possibility of using meaningful personality vari-
ables as a basis for assigning students to classes.

The expected difference in achievement between
externals and internals did not materialize. On the
basis of the available data, this can best be accounted
for by the exceptionally high level of achievement by
the externals in the role-assumption treatment. These
data indicate that, under the ordinary circumstances
of a structured treatment, the difference in achieve-
ment between externals and internals would have
been observed.

The hypothesized difference in achievement be-
tween students assigned to the role-assumption treat-
ment and students assigned to the structured treat-

ment was supported. This finding can be accounted
for primarily by the differential achievement of ex-
ternals in the two treatments. The evidence clearly
indicates that it would be best to employ the role-as-
sumption treatment regardless of whether the stu-
dents are classified as externals or internals. Since
there was no difference between treatments for in-
ternal students, but a large difference between treat-
ments for external students favoring the role-assump-
tion treatment, the decision of which method to use
has to favor the role-assumption treatment. These
findings indicate that a knowledge of the type of stu-
dents involved can be an important variable to con-
sider. If a large proportion of the students are exter-
nals, as is the case at the North Carolina Advance-
ment School, then it would be of particular importance
to implement a role-assumption treatment.

These results provide some support for the points
of view expressed by Thelen (16, 17) and Siegel (13).
Thelen has placed emphasis upon the development of
process skills and the use of inquiry methods to de-
velop stable and transferable knowledge. The role-
assumption treatment used in this research may be
broadly classified as a type of inquiry approach. Sie-
gel has recently emphasized that teachers must shift
from didactic teaching methods to the more complex
dialectic teaching processes. The results of this re-
search support this concept.

REFERENCES

1. Bettelheim, Bruno, " Autonomy and Inner Free-
dom: Skills of Emotional Management,"in Rub-
en, Louis J(ed.) Life Skills in School and Society,
Association for Supervision and Curriculum De-
velopment, Washington, D. C. , 1969, pp. 73-94.

2. Crandall, V. C. ; Katkovsky, W. ; Crandall, V. J. ,
" Children's Beliefs in Their Own Control of Re-
inforcements in Intellectual-Academic Achieve-
ment Situations," Child Development, 36:91-109,
1965.

3. Crandall, V. J. ; Katkovsky, W. ; Preston, A. ,"Mo-
tivational and Ability Determinants of Young Chil-
dren's Intellectual Achievement Behaviors,"
Child Development, 33: 643-661, 1962.

4. Cronbach, L. J. , " The Two Disciplines of Scien-
tific Psychology," American Psychologist, 12:
671-684, 1957.

5. Educational Testing Service (ETS), Sequential
Tests of Educational Progress (STEP): Science,
Cooperative Test Division, ETS, Princeton, New
Jersey, 1957.

6. Hullfish, H. Gordon; Smith, Philip G. , Reflective
Thinking: The Method of Education, Dodd, Meade
and Company, New York, 1963.

7. Jackson, P. W. , " School Achievement and Alien-
ation: Notes on the Psychology of Classroom Fail-
ure," in Landy, Edward; Kroll, Arthur,(eds.)Guid-
ance in American Education II: Current Issues
and Suggested Action, Harvard University Press,
Cambridge, Massachusetts, 1965, pp. 141-154.

8. Lefcourt, H. M. , " Internal Versus External Control of Reinforcement: A Review," Psychological Review, 65: 206-220, 1966.

9. Maslow, Abraham, " Deficiency Motivation and Growth Motivation, " in Jones, Marshall R. (ed.) Nebraska Symposium on Motivation, University of Nebraska Press, Lincoln Nebraska, 1955, pp. 1-30

10. Rogers, Carl R. , " The Actualizing Tendencies in Relation to Motives and to Consciousness," in Jones, Marshall R. , (ed.) Nebraska Symposium on Motivation, University of Nebraska Press, Lincoln Nebraska, 1963, pp. 1-24.

11. Rotter, J. B. , " Generalized Expectancies for Internal Versus External Control of Reinforcement, " Psychological Monographs, 80: (1, number 609), 1966.

12. Rotter, J. B. , Social Learning and Clinical Psychology, Prentice-Hall, Englewood Cliffs, New Jersey, 1954.

13. Siegel, L. (ed.), Instruction: S o m e Contemporary Viewpoints, Chandler, San Francisco, California, 1967.

14. Slosson, R. L. , The Slosson Intelligence T e s t (SIT) for Children and Adults, Slosson Educational Publications, East Aurora, New York, 1963.

15. Thelen, H. A. , Classroom Grouping for Teachability, Wiley, New York, 1967.

16. Thelen, H. A. , Dynamics of Groups at Work, University of Chicago Press, Chicago, 1954.

17. Thelen, H. A. , Education and the Human Quest, Harper Brothers, New York, 1960.

18. Watson, F. G. , "Research on Teaching Science," in Gage, N. L. (ed.), Handbook of Research on Teaching, Rand McNally, Chicago, 1963.

A REGRESSION APPROACH

TO EXPERIMENTAL DESIGN

JOHN D. WILLIAMS
The University of North Dakota

ABSTRACT

This paper describes a regression approach to experimental design. Typically, the educational research-er is familiar with the usual analysis of variance techniques and is unaware of the versatility of multiple regres-sion for problem solving. The specific approach used in this paper assumes only a familiarity with experimental design and access to a general purpose multiple regression program which is likely to be on hand at any comput-er installation. Examples of the t-test, 1-way analysis of variance, and the treatment X subjects design from a regression standpoint are given.

MANY EDUCATIONAL researchers will have had an exposure to several applications of anal-ysis of variance, and quite commonly this will have occurred in a course called "experimental design" or some similar title. The classical analysis of vari-ance approach will commonly expose the student to several specialized topics, such as the treatment X subjects design, the factorial design, the randomiz-ed blocks design, the split-plot design, and the anal-ysis of covariance.

It is the contention of this writer that researchers who have received their statistical training in the classical approach have at least two difficulties that limit their data analysis procedures. For any large amount of data, a computer would seem a necessary adjunct to their analysis; however, there may be a strong tendency for researchers trained in the clas-sical approach to rely heavily on stock programs when using the computer. Further, they are likely to look for a different stock program for each design. A second difficulty that may be likely to occur is that a person trained in the classical approach will try to make his design fit into an existing design rather than formulate the model most appropriate for his unique problem.

The present discussion focuses on some designs familiar to the educational researcher. The usual analysis of variance presentation occurs in such text-books as Lindquist (7) or Edwards (4). While this presentation is concerned only with well-known de-signs, it is hoped that the reader will be able to at least start to formulate his research problems so that he might be able to submit his data to a comput-er and receive answers to specific research ques-tions that are of interest.

A word is in order as to the direction of the dis-cussion—only one computer program, the general purpose multiple regression program likely to be on hand at any modern computer installation, will be considered. Thus, a minimum of sophistication is necessary regarding programming. A somewhat similar approach is used by Kelley and others (6), who employ programs known as LINEAR and DATRAN. Bottenberg and Ward (2) have used a similar ap-proach (also using DATRAN) in presessions of the American Educational Research Association annual conventions.

A first step toward using multiple regression as a problem-solving technique would be to formulate models for multiple regression for several of the more familiar designs. A natural starting point is the t-test.

THE t-TEST

Consider the following data:

Group 1	Group 2
25	25
24	23
23	21

t test data continued from previous page.

Group 1	Group 2
23	17
22	22
21	20
18	15
17	14
16	12
16	10
16	10
15	12
14	11
13	10
13	7
13	6
12	6
11	8
10	4
9	7

When the usual t-test is run, a value of 1.97 is found. Using a regression approach, it is first useful to define two binary predictors, X_1 and X_2:

X_1 = 1 if the score is from a member of Group 1; and 0 otherwise,

X_2 = 1 if the score is from a member of Group 2; and 0 otherwise.

A linear model can be written for this situation:

$Y = b_0 + b_1X_1 + b_2X_2 + e$

where:
Y = the criterion score,

b_0 = the Y-intercept,

b_1 = the regression coefficient for X_1,

b_2 = the regression coefficient for X_2,

e = the error involved in prediction.

The preceding information can be put in a table (see Table 1) the format of which is particularly helpful for preparation of the data cards for submission to the computer. One card can be prepared for each of the observed scores, together with the information regarding group membership.

The conceptualizing that usually takes place during the use of the t-test might center around the question, do the means differ significantly? On the other hand, in a regression formulation, the researcher's thought process may involve using the knowledge of group membership to predict the criterion scores. In the final analysis, both approaches use the same linear model.

If the multiple regression program is used with the data from Table 1, with the two predictor variables, X_1 and X_2, and with the Y variable as the criterion, the probable result is that the program will not run and will simply report back something like "MATRIX SINGULAR SELECTION IS SKIPPED AS REQUESTED." For the person unfamiliar with the computer, many different reactions can occur,

from disgust to "I knew it couldn't do it." Without going into technical details, the problem is one of supplying too much information. Going back to one of the predictor variables, say X_1, either a 1 or a 0 is recorded for every criterion measure. The next column, X_2, is simply the reverse of X_1, that is, if there was a 0 in X_1, then X_2 has to be a 1.

Thus, only one of the predictor variables is necessary to impart all the information regarding group membership.

TABLE 1

REGRESSION FORMULATION OF THE t-TEST

Y	X_1	X_2
25	1	0
24	1	0
23	1	0
23	1	0
22	1	0
21	1	0
18	1	0
17	1	0
16	1	0
16	1	0
16	1	0
15	1	0
14	1	0
13	1	0
13	1	0
13	1	0
12	1	0
11	1	0
10	1	0
9	1	0
25	0	1
23	0	1
21	0	1
17	0	1
22	0	1
20	0	1
15	0	1
14	0	1
12	0	1
10	0	1
10	0	1
12	0	1
11	0	1
10	0	1
7	0	1
6	0	1
6	0	1
8	0	1
4	0	1
7	0	1

For the information in Table 1, X_1 was used as the predictor of the criterion variable, using the general purpose multiple regression program.

TABLE 2

MULTIPLE REGRESSION SOLUTION FOR t-TEST SITUATION

SELECTION 1

Variable No.	Mean	SD	Correlation X vs Y	Regression Coefficient	Standard Error of Regression Coefficient	Computed t Value	Beta
1	0.50000	0.50637	0.30497	3.55000	1.79835	1.97403	0.30497
Dependent							
3	14.77500	5.89431					
Intercept		13.00000					
Multiple Correlation		0.30497					
Standard Error of Estimate		5.68689					

ANALYSIS OF VARIANCE FOR THE REGRESSION

Source of Variation	Degrees of Freedom	Sum of Squares	Mean Squares	F Value
Attributable to regression	1	126.02502	126.02502	3.89678
Deviation from regression	38	1228.94995	32.34077	
Total	39	1354.97485		

The intercept is also included. This value is 13.00 and is the mean of the group that was coded as 0 on the X_1 variable. The multiple correlation, R, is .304097. Since there is only one predictor variable, this value is in fact a 0-order correlation. Both the standard error of estimate and an analysis of variance for the regression are included in the printout. An option is available in the multiple regression program for a table of residuals if they are desired. Table 2 contains the printout (not including the table of residuals) for the data in Table 1. The regression coefficient is for the predictor variable X_1; we have, in effect set $b_2 = 0$ by this approach.

It can be noted that the results that were obtained by the usual t-test were also found by using the regression approach. In addition to the computed t value ($t = 1.97403$), the F value of 3.897 is also equal to t^2. Additionally, if the correlation coefficient is squared ($r^2 = .0925$), this indicates that 9.25 percent of the criterion variance can be accounted for by knowledge of group membership.

1-WAY ANALYSIS OF VARIANCE

The situation for analysis of variance can be seen to be very similar to the regression analysis for the t-test. Consider the following data:

Group 1	Group 2	Group 3	Group 4
4	13	11	10
2	10	9	9
0	7	7	8

The usual analysis of variance for this data is reported in Table 3.

TABLE 3

SUMMARY TABLE FOR THE ANALYSIS OF VARIANCE

Source of Variation	Degrees of Freedom	Sum of Squares	Mean Squares	F
Among groups	3	123.00	41.00	9.11
Within groups	8	36.00	4.50	
Total	11	159.00		

The intent here is to show that the same results can also be accomplished using the multiple regression program, rather than focusing on the analysis of variance per se.

To conceptualize the problem for a regression analysis, it is helpful to define a set of binary predictors. Four binary predictors can be defined to correspond to the four groups:

X_1 = 1 if the score is from a member of Group 1; 0 otherwise,

X_2 = 1 if the score is from a member of Group 2; 0 otherwise,

X_3 = 1 if the score is from a member of Group 3; 0 otherwise,

X_4 = 1 if the score is from a member of Group 4; 0 otherwise.

A linear model can be written for this situation:

$$Y = b_0 + b_1X_1 + b_2X_2 + b_3X_3 + b_4X_4 + e$$

where:

Y = the criterion score

b_0 = the Y-intercept

b_1 = the regression coefficient for X_1

b_2 = the regression coefficient for X_2

b_3 = the regression coefficient for X_3

b_4 = the regression coefficient for X_4

e = the error involved in prediction.

The information regarding group membership and the criterion scores can be put into a format similar to Table 1 and is found in Table 4.

TABLE 4

REGRESSION FORMULATION FOR 1-WAY ANALYSIS OF VARIANCE

Y	X_1	X_2	X_3	X_4
4	1	0	0	0
2	1	0	0	0
0	1	0	0	0
13	0	1	0	0
10	0	1	0	0
7	0	1	0	0
11	0	0	1	0
9	0	0	1	0
7	0	0	1	0
10	0	0	0	1
9	0	0	0	1
8	0	0	0	1

Again, it can be remembered that the last column is actually not adding any new information and can be considered redundant. If b_4 is set equal to 0, the prediction equation can be rewritten as

$$Y = b_0 + b_1X_1 + b_2X_2 + b_3X_3 + e$$

Using this setup, the previous data was analyzed with the general purpose multiple regression program.

Again, the printout will normally include the means, standard deviations, correlations with the criterion, regression coefficients, standard errors of the regression coefficients, computed t values, and the beta coefficients. The intercept for this data is given as 9.00. The multiple correlation, R, is equal to .87954, and the standard error of estimate is 2.1213. In the analysis of variance for regression, the following summary table (Table 5) is routinely printed out (there is also included in the printout the previously mentioned items, but they are omitted here).

TABLE 5

ANALYSIS OF VARIANCE FOR REGRESSION

Source of Variation	Degrees of Freedom	Sum of Squares	Mean Squares	F
Attributable to Regression	3	123.00	41.00	9.111
Deviation from Regression	8	36.00	4.50	
Total	11	159.00		

The usage of the regression program should become apparent. Not only is the data available that normally is a part of analysis of variance, but also, a measure of the amount of variance that can be accounted for by the predictor variables can be found. Here R = .87954 and, therefore, R^2 = .7755; thus 77.55 percent of the criterion variance can be accounted for by knowledge of the group membership.

One additional advantage is pedagogical. As the student becomes familiar with the regression approach, he can more fully understand the logic of both regression and experimental analysis.

FACTORIAL DESIGNS

The factorial design (called such by Lindquist), or fixed effects design, can be first viewed as a 2-way analysis of variance. Three-way and higher dimensional designs follow the same conceptualization. The basic difference between the 1-way and 2-way designs, in addition to the additional variables (factors), is the inclusion of the interaction effect. Jennings[5] has discussed in some detail formation of the 2-way fixed effects analysis conceptualized in the regression approach. Rather than duplicate his effort, the

reader is encouraged to read Jenning's original article.

TREATMENT X SUBJECTS DESIGN

The treatment X subjects design is described in detail in Lindquist (7), and is a repeated measures design wherein each S serves as his own control. Consider the following data:

Subject	Treatment 1	Treatment 2	Treatment 3
1	18	27	15
2	17	24	14
3	14	13	12
4	5	8	6
5	11	14	10
6	9	12	8
7	14	16	15
8	12	17	9
9	22	21	16
10	10	18	15

If the usual analysis of variance is performed, a summary table (see Table 6) is found.

TABLE 6

ANALYSIS OF VARIANCE FOR TREATMENT X SUBJECTS DESIGN

Source of Variation	Degrees of Freedom	Sum of Squares	Mean Squares	F
Treatments	2	136. 27	68. 13	11. 52
Subjects	9	521. 20	57. 91	
Error	18	106. 39	5. 91	
Total	29	763. 86		

On the other hand, if the problem is viewed from a regression approach, a solution can be obtained in stages. First, a Full Model is to be constructed. Very simply, a full model will contain all of the information available for a given situation. For the treatment X subjects design, two bits of information are available for each criterion score, the treatment involved, and the subject involved. If a set of thirteen binary predictors are constructed, the full model will become

$$Y = b_0 + b_1X_1 + b_2X_2 + b_3X_3 + b_4X_4 + b_5X_5 +$$
$$b_6X_6 + b_7X_7 + b_8X_8 + b_9X_9 + b_{10}X_{10} + b_{11}X_{11} +$$
$$b_{12}X_{12} + b_{13}X_{13} + e_1$$

where:

Y = the criterion score

X_1 = 1 if the score is from subject number 1;

0 otherwise,

X_2 through X_{10} are defined similar to X_1,

X_{11} = 1 if the score is from a member of treatment 1; 0 otherwise,

X_{12} = 1 if the score is from a member of treatment 2; 0 otherwise,

X_{13} = 1 if the score is from a member of treatment 3; 0 otherwise,

e_1 = the error in prediction.

The regression coefficients are defined in terms of the corresponding predictor, with b_0 being the Y-intercept. The information from the preceding can be conveniently put into a table (see Table 7) so that it can be readily put on IBM cards.

As happened before, we actually have too much information. That is, X_{10} is determined by the previous nine predictors. Thus, it can simply be discarded as a predictor. This is the same as $b_{10} = 0$. Likewise, X_{13} is dependent upon X_{11} and X_{12}, and can also be discarded as a predictor. This also is the same as setting $b_{13} = 0$. The full model will then use eleven predictors and be referred to as selection one. It will also be called Model I.

Two restricted models will be defined. First, ten binary predictors can be constructed, one for each subject, and then the last predictor is deleted, the regression equation becomes:

$$Y = b_{14} + b_{15}X_1 + b_{16}X_2 + b_{17}X_3 + b_{18}X_4 + b_{19}X_5 +$$
$$b_{20}X_6 + b_{21}X_7 + b_{22}X_8 + b_{23}X_9 + e_2$$

where:

Y = the criterion variable,

X_1 through X_9 are defined the same as in the Full Model (Model I),

b_{14} = the Y-intercept,

b_{15} = the regression coefficient corresponding to X_1,

b_{16} through b_{23} are similarly defined,

e_2 = the error in prediction for subjects.

This formulation can be called Model II and will yield the subjects effect.

Finally, a model (Model III) can be defined for treatments (columns) and is

$$Y = b_{24} + b_{25}X_{11} + b_{26}X_{12} + e_3$$

where:

Y = the criterion score,

TABLE 7

REGRESSION FORMULATION FOR TREATMENT X SUBJECTS DESIGN

Y	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}	X_{11}	X_{12}	X_{13}
18	1	0	0	0	0	0	0	0	0	0	1	0	0
27	1	0	0	0	0	0	0	0	0	0	0	1	0
15	1	0	0	0	0	0	0	0	0	0	0	0	1
17	0	1	0	0	0	0	0	0	0	0	1	0	0
24	0	1	0	0	0	0	0	0	0	0	0	1	0
14	0	1	0	0	0	0	0	0	0	0	0	0	1
14	0	0	1	0	0	0	0	0	0	0	1	0	0
13	0	0	1	0	0	0	0	0	0	0	0	1	0
12	0	0	1	0	0	0	0	0	0	0	0	0	1
5	0	0	0	1	0	0	0	0	0	0	1	0	0
8	0	0	0	1	0	0	0	0	0	0	0	1	0
6	0	0	0	1	0	0	0	0	0	0	0	0	1
11	0	0	0	0	1	0	0	0	0	0	1	0	0
14	0	0	0	0	1	0	0	0	0	0	0	1	0
10	0	0	0	0	1	0	0	0	0	0	0	0	1
9	0	0	0	0	0	1	0	0	0	0	1	0	0
12	0	0	0	0	0	1	0	0	0	0	0	1	0
8	0	0	0	0	0	1	0	0	0	0	0	0	1
14	0	0	0	0	0	0	1	0	0	0	1	0	0
16	0	0	0	0	0	0	1	0	0	0	0	1	0
15	0	0	0	0	0	0	1	0	0	0	0	0	1
12	0	0	0	0	0	0	0	1	0	0	1	0	0
17	0	0	0	0	0	0	0	1	0	0	0	1	0
9	0	0	0	0	0	0	0	1	0	0	0	0	1
22	0	0	0	0	0	0	0	0	1	0	1	0	0
21	0	0	0	0	0	0	0	0	1	0	0	1	0
16	0	0	0	0	0	0	0	0	1	0	0	0	1
10	0	0	0	0	0	0	0	0	0	1	1	0	0
18	0	0	0	0	0	0	0	0	0	1	0	1	0
15	0	0	0	0	0	0	0	0	0	1	0	0	1

X_{11} and X_{12} are defined as before,

b_{24} = the Y-intercept for Model III,

b_{25} = the regression coefficient for X_{11},

b_{26} = the regression coefficient for X_{12},

e_3 = the error in prediction for treatments.

The variable X_{13} was deleted here for the same reason that it was deleted in the Full Model; that is, it introduces no new information, and is thus redundant. Model III will be called selection three.

Using three selections, each of the sources of variation can be separately determined. After combining the results of the three selections, essentially the same results as found in Table 7 can be gathered. Actually, any one of the selections could have been omitted and the other portion could have been found as a residual.

Using the multiple correlations from each selection, the following additional results were found:

(a) With the full model, R = .92774, and R^2 = .8607.

(b) With the 9-predictor model system, that is, the "subject" effect, R = .82602, and R^2 = .6823.

(c) With the regression model for treatment effects, R = .42236, and R^2 = .1784.

Hence, it can be seen that the "subjects" accounted for the most variance.

While this entire analysis has been concerned with duplicating the treatment X subjects design described by Lindquist, exactly the same set of linear models could be written for the randomized block design.

OTHER DESIGNS

Other common designs such as Latin squares and the split-plot designs can also be conceptualized and completed using a regression approach. Schmid (9) has demonstrated by an example that the analysis of covariance can be formulated from a regression approach and he has shown that the two approaches yield an identical F value. Because of the simplicity of the regression approach to analysis of covariance, it is worth further discussion here. Also, it should be pointed out that the name "analysis of covariance" may not specifically be used by an author using the regression approach. Bottenberg and Ward (2) use the term "concomitant variable."

The analysis of covariance can be accomplished by successive uses of the general purpose multiple regression program. Suppose, for example, that the researcher wishes to use two covariates. Let these predictors be notated X_1 and X_2. Suppose also that three groups are being used as the treatment groups. The full model can be given as:

$$Y = b_0 + b_1X_1 + b_2X_2 + b_3X_3 + e_1$$

where:

Y = the criterion score,

X_1 = the first covariate,

X_2 = the second variate,

X_3 = 1 if the score is from a member of Group 1; 0 otherwise,

X_4 = 1 if the score is from a member of Group 2; 0 otherwise,

e_1 = the error involved in prediction for this model (the Full Model),

b_0 = the Y-intercept for the Full Model,

b_1 through b_4 are regression coefficients for their respective predictor variables.

It should be noticed that the third group has simply been identified by not having membership in groups 1 and 2, as was done in the regression conceptualization of 1-way analysis of variance. With the covariance design, it is useful to utilize the R^2 value directly from the printout of the multiple regression program.

A restricted model can be formed, using only the covariates:

$$Y = b_5 + b_6X_1 + b_7X_2 + e_2$$

Then the test of significance between the full and restricted models can be completed:

$$F = \frac{(R^2_{FM} - R^2_{RM})/df_n}{(1 - R^2_{FM})/df_d}$$

The R^2_{FM} is a symbol for the R^2 term from the full model, and R^2_{RM} is a symbol for the R^2 term from the restricted model. The degrees of freedom in the numerator (df_n) is K-1 where K is the number of experimental groups. The degrees of freedom in the denominator (df_d) is N-C-K where N is the number of subjects, C is the number of covariates, and K is the number of experimental groups. It should be pointed out that this formulation does not provide a test for interaction.

Actually, almost any existing analysis of variance design can be conceptualized and completed by using the general purpose multiple regression program. If the regression approach had nothing else to offer, the ease of using the computer as an adjunct to problem solving would be worthwhile in itself.

However, a most important point needs to be considered with experimental design and, especially, formal courses in that subject. A potential student might rightly assume that one expected behavioral outcome of his having taken a course in experimental design is that he can design an experiment. If comments such as Addleman's (1) are a true reflection of the actual status of that course, the student instead learns about several well-known existing

designs but does not actually reach the stage of flexibility to design his own research.

While it has by no means been demonstrated here, once the researcher becomes familiar with the use of the regression program, he can take at least one step toward his own experimental design. For example, suppose a researcher were interested, for research purposes, in a curvilinear interaction. If he were to follow the routine of searching existing programs for such a situation, he may well change the course of his experimentation so that he could use a more familiar design. If, on the other hand, the researcher was familiar with both the linear models and the general purpose multiple regression program, he could pursue the question of interest.

It should again be emphasized that the content of this presentation has assumed only two things: the reader has a background in experimental design and has access to a general purpose multiple regression program likely to be found at any modern computer installation. As indicated earlier, other approaches to problem solving by multiple regression also use other subroutines. It would be useful, however, for the researcher to try the multiple regression program already available to him. The versatility of the program can soon become apparent.

Two articles of related interest would be useful in pursuing the multiple regression approach. Cohen (3) gives some insight into the difficulty researchers have had realizing the use of multiple regression for problem solving. Ward (10) discusses in some detail, four different approaches that researchers with different statistical backgrounds would use as they relate to experimental design. He also argues effectively for the usage of multiple regression as a method of problem solving.

Assuming the reader has attempted the use of the general purpose multiple regression program, he might ask, "What is next?" The two articles mentioned in the previous paragraph are a good starting point in the use of the regression approach to problem solving. Depending on the reader's interest and background, several other suggestions might be made. It would be worthwhile to consider coursework designed to help research workers formulate models. Besides the sources already given, another excellent text is one by Mendenhall (8). While it would be fair to say that the points of view of the various authors in the sources listed are not identical, they retain the same flavor of using regression and linear models for the solution of research questions.

SUMMARY

The intention of this presentation has been to introduce the educational researcher who typically has a background in experimental design to the regression approach to problem solving. With the increasing availability of computers, the researcher can have an expectancy of having much more flexibility in the design of his own research.

With the availability of several stock programs, the researcher may not be aware of the versatility of the multiple regression program. Several of the more common types of designs have been examined from a regression approach. This has been done so

that the researcher might become convinced that this approach will include the designs he is familiar with, but from a different viewpoint. The most important contribution occurs after the researcher feels at ease with the multiple regression program. Then he can truthfully design his own research rather than try to fit some existing design.

This presentation has concerned itself only with the general purpose multiple regression program which is likely to be available at any modern computer installation. Becoming familiar with this program might be considered a first step toward more flexibility for the individual researcher in his experimental design.

REFERENCES

1. Addleman, S., "Book Review of Experimental Design: Procedures for the Behavioral Sciences," Journal of the American Statistical Association, 64:1096-1097, 1969.

2. Bottenberg, R. A.; Ward, J. H., Applied Multiple Linear Regression, Personnel Research Laboratory, Aerospace Medical Division, (PRL-TDR-63-6) Lackland Air Force Base, Texas, 1963.

3. Cohen, J., "Multiple Regression as a General Data-Analytic System," Psychological Bulletin, 70:426-443, 1968.

4. Edwards, A. L., Experimental Design in Psychological Research, Third Edition, Holt, Rinehart, and Winston, New York, 1968.

5. Jennings, E., "Fixed Effects Analysis of Variance by Regression Analysis," Multivariate Behavioral Research, 2:95-108, 1967.

6. Kelley, F. J.; Beggs, D. L.; McNeil, K. A.; Eichelberger, T., Multiple Regression Approach, Southern Illinois University Press, Carbondale, Illinois, 1969.

7. Lindquist, E. F., Design and Analysis of Experiments in Psychology and Education, Houghton Mifflin, Boston, Massachusetts, 1953.

8. Mendenhall, W., Introduction to Linear Models and the Design and Analysis of Experiments, Wadsworth Publishing Company, Belmont, California, 1968.

9. Schmid, J., "Research Design with Multiple Regression," dittoed, Colorado State College, Greeley, 1967.

10. Ward, J. H., "Synthesizing Regression Models, An Aid to Learning Effective Problem Analysis," The American Statistician, 23:14-20, April 1969.

EXPERIENCE, SKILL, EXPRESSED FEAR, AND

EMOTIONAL REACTION TO MOTOR SKILLS

PERFORMED UNDER CONDITIONS OF HEIGHT

WANEEN WYRICK
The University of Texas at Austin

ABSTRACT

This study determined the extent to which Ss, grouped on the basis of excellent, average or poor on two motor tasks performed 4 feet above the floor, could be differentiated by their previous pleasant and unpleasant experiences with height, expressed fear of height, emotional response to height, risk-taking activity level, running skill, and cross-step skill. Six trials of running and cross-stepping on a balance beam, three trials of these tests raised 4 feet above the floor, and a self-report inventory were completed by 139 college women. A multiple discriminant analysis indicated that the criterion groups were differentiated on both tasks by their motor skill and their current risk-taking activity level, but not by their past experiences or expressed fear of heights.

EDUCATORS generally assume that pleasant past experiences in an activity and the attainment of a moderate level of technique are two factors that may positively influence an individual's willingness to attempt new skills. Conversely, it is thought that when students come to the learning experience with a history of failures resulting in unpleasant experiences, the student may be fearful and quite hesitant to participate in the activity. This is vividly true in physical education classes, where many of the gross motor skills that are taught must be performed at least partially successfully during the first attempt in order to prevent injury. This is particularly true in those activities which are performed at a height. Some students, more than others, seem to be preoccupied with the possibility that poorly coordinated attempts at springboard diving skills, gymnastic maneuvers, rope climbing, and trampoline stunts may result in physically injurious and painful falls. It is not uncommon for the physical educator to note that a student's coordinated attempts — which were quite satisfactory when executed on the floor in a learning progression — degenerate rapidly when the skill is attempted above the floor.

The detrimental effect that fear or anxiety has upon motor performance is a very real phenomenon which has been identified by Wherry (4, 5) as antic-

ipatory physical threat stress (APTS). Moderate levels of APTS apparently enhance performance, while unusually high levels result in disrupting behavior. Complex responses probably deteriorate under APTS more than simple responses. When Ss are provided with an opportunity to avoid physical threat by skilled motor performance, their performance is enhanced (2).

The levels of APTS generated are described as a function of an individual's past experience, which in turn patterns his perception of the probability of physical threat, the proximity of physical threat, and the unpleasantness of the event. A factor that apparently generates substantial levels of APTS in some students is the requirement that a motor task be performed at a substantial height. The fact that there is considerable variation in students' emotional response to a potentially dangerous task evokes an interesting question. Why are some students terrified at the thought of moving their bodies through space high above the floor, and other students see it only as an interesting task? What combination of variables might enable a teacher to predict performance that will degenerate under APTS ?

It is tempting to hypothesize that an individual who has had pleasurable childhood experiences with height, such as playing in swings, jungle gyms, and

tree houses, perceives height as adding to the enjoyment or thrill of the motor task. It is also tempting to suggest that if in addition to these pleasant experiences an individual visualizes himself as skillful in motor activities, he is likely to anticipate the event of falling to be of low probability. Another possibility is that a person who has been accustomed to playing at heights and perhaps falling occasionally, may not perceive the event of falling as having the same degree of unpleasantness as that person who was protected and who currently rarely places himself in unstable positions above the floor.

It would be quite beneficial to know whether knowledge of skill level and past experiences with heights could enable the physical educator to predict initial and tentative early performances that will deteriorate when performed at a height. If past experience and skill level were predictive, injuries could be avoided. In addition, this concept would certainly be considered by the constructors of elementary physical education curricula.

It would also be helpful to know whether knowledge of a student's self-reported emotional or physiological reaction to height as well as his current self-initiated risk-taking activity level would provide the educator with information that would predict early successes or failures. Although many physiological responses have been categorized as indicators of emotional arousal (3), it is not known whether these responses are predictors of behavior under APTS.

Human behavior can no longer be explained in terms of single variables as well as it can in terms of multiple variables interpreted in relationship to each other. The specific contribution of experience, skill, and emotional arousal to motor performance under height conditions is not the question of import, but rather, how the variables of past pleasant and unpleasant experiences, motor skill in the task, emotional response to the height, and current risk-taking activities interact to determine individuals' initial performances of a motor skill high above the floor. Finally, does the complexity of the motor task differentially affect the use of these descriptive variables in predicting performance under task-height conditions?

PROBLEM

This study determined the extent and manner in which Ss, grouped on the basis of excellent, average, or poor on two dynamic gross motor tasks performed 4 feet above the floor, may be differentiated by the following set of descriptive variables operating together: pleasant experiences with height, previous unpleasant experiences with height, expressed fear of height, emotional response to height, risk-taking activity level, running skill, and cross-stepping skill.

PROCEDURES

Two basic types of data were obtained: the criterion scores that reflected Ss' ability to perform a motor task 4 feet above the floor, and the descriptive scores, that placed the S on a continuum for each of seven variables.

Criterion Scores

College freshmen and sophomore women (N=139) were the Ss of this investigation, and completed three trials at each of two gross motor tasks. Both tasks involved traveling the length of a 4-foot high balance beam as quickly as possible. Subjects began by standing on the end of the beam in such a way that one foot depressed a microswitch. When the S began the task by lifting her foot from the microswitch, a chronoscope was activated and continued timing until a microswitch at the opposite end of the beam was depressed by the S. The two tasks were (a) running the length of the beam, and (b) traversing the beam with a cross-step sideward run. The within-day reliability coefficients were .94 and .86, respectively. Excellent, average, and poor groups were delineated on the basis of performance on each of these tasks.

Descriptive Variables

Pleasurable and unpleasant experiences that Ss had had with heights, their expressed fear of high places, and their current level of participation in risk-taking activities were estimated by sub-scores on a self-report inventory. Emotional responses to height were also obtained by a self-report inventory which was administered immediately after Ss had traversed the beam for the first time at its regulation height of 4 feet. In responding to the emotional response inventory, Ss placed themselves on a 5-point scale for each of fifteen physiological autonomic nervous system responses commonly reported as accompanying fear or stress anticipation.

Motor skill was defined as the performance level of the S on the running and cross-step tasks without the height factor. The beam was placed flat on the floor and Ss were timed throughout three trials on both tasks. Within-day reliability coefficients were .90 and .87, respectively. Scoring and timing procedures were identical to those used throughout the criterion tasks.

Testing Schedule

Subjects first became familiar with the skill tasks and laboratory equipment by completing, individually under experimental conditions, three trials on the running tasks and the cross-step task. These data were not utilized, since the trials were undertaken primarily to familiarize the S and to stabilize the drastic reduction of time between trials one and two that occurs as a result of learning (6, 7).

One week later, the motor skill scores were obtained by procedures that were identical to those used in the familiarization testing session. The self-report inventory scores were obtained during group testing sessions throughout the third week. Subjects' emotional responses to height were obtained during the fifth week when Ss, for the first time, walked across the balance beam at its full height. Upon descending, Ss completed the emotional response inventory. The high performance criterion scores were obtained during the sixth week. Subjects completed three trials of the running task and three trials of the cross-step task under conditions identical to those of the familiarization and skill testing sessions. All Ss were individually tested by the same investigators, and prior to this last testing session, Ss had no knowledge that they would be asked to perform these skill tasks at a height of four feet.

ANALYSIS

All Ss were described, on the basis of the data collected throughout the 6-week period, in terms of their position on a continuum for each of seven variables that were postulated a priori as variables that might singly or in combinations describe their motor performance high above the floor. The high running and cross-step task scores, serving as the criterion score representing Ss' ability to perform under the stress of height, were used to divide the Ss into three criterion groups: excellent (Q_1), average ($Q_{2,3}$), and poor (Q_4) performers. A multiple discriminant analysis was used to determine the extent and manner in which the descriptive variables might also describe the category in which Ss were classed when performing under the stress of height (1).

The discriminant analysis used in this study may be conceptualized as an extension of a single-classification analysis of variance which includes simultaneously several variables in order to determine whether significant group differences exist. The technique produces discriminant functions that include the original variables that discriminate the groups; discriminant weights, which are the relative contribution of each variable to the discriminant function; a Wilkes Lambda, indicating the probability level for the null hypothesis of equality of group centroids; and a percent of variance extracted, indicating the percent of total discriminating power of the test battery contained in each discriminant function.

RESULTS

High Running Task

As may be seen in Table 1, the excellent, average, and poor groups as categorized on the high running task were also differentiated by a function which includes their emotional responses to height, their current risk—taking activity level, their running skill, and their cross step skill. These descriptive variables, in terms of their ability to categorize the three criterion groups, combined to be highly significant predictors which accounted for almost 92 percent of the total variance. They were, therefore, substantially more important in terms of Ss' performance at a height than those variables of past experiences—either pleasant or unpleasant—and the Ss' expressed fear of high places. Although pleasant experiences with heights and expressed fear of high places combined to form a discriminant function, it was not a significant function.

The criterion group means were distributed in a consistently linear pattern on those four descriptive variables which combined to discriminate among the three criterion groups (Table 2). The excellent performers at heights reported significantly fewer emotional responses, such as tachycardia, cold palms, profuse perspiration, nausea, than the average group, while they in turn reported fewer than the poor group. The excellent group reported that they participated in more risk-taking activities, such as water skiing, riding, diving, skydiving, and ski jumping, than the other two groups. The excellent group was also more skillful on the low running task than the average group, while that group in turn was superior to the poor group. The three groups did not appear to have different past experiences with heights, as is indicated by the almost identical group means in Table 2. This finding

TABLE 1

DISCRIMINANT AXES AND COMPONENTS FOR GROUPS CATEGORIZED ON THE HIGH RUNNING AND HIGH CROSS-STEP TASK

Components	High Running Discriminant Weights [a]	High Cross-stepping Discriminant Weights [b]
Discriminant Axis I		
Emotional Response to Height	.63	.35
Risk-taking Activity level	-.49	-.46
Running Skill	.84	.68
Cross-step Skill	.60	.91
Discriminant Axis II		
Pleasant Experience with High Places	.76	.50
Expressed Fear	.36	.59

[a] Total trace = 100 % λ = .463 $F_{14, 260}$ = 7.39

P = > .001

Axis I = 94.90% of Variance. x^2 = 83.71, df = 8, P = > .001

Axis II = 5.10 % of Variance. x^2 = 6.10, df = 6 P = .410 NS

[b] Total trace = 100 % λ = .465. $F_{14, 260}$ = 8.67

P = > .001

Axis I = 98.33 % of Variance. x^2 =100.14 df = 8 P = > .001

Axis II = 1.67 % of Variance. x^2 = 2.51 df = 6 P = .870 NS

may be illustrated by the fact that, of those Ss who reported they had broken a bone as a result of a fall, sixteen were in the excellent group, eight were in the average group, and eleven were in the poor group. There apparently was no pattern of differentiation in unfortunate or unpleasant experiences in the childhood of the Ss among the three groups. The expressed fear of height was essentially the same among the three groups also.

TABLE 2

MEANS, UNIVARIATE F RATIOS, AND PROBABILITY LEVELS FOR GROUPS CLASSIFIED BY PERFORMANC
ON HIGH RUNNING TASK

Descriptive variables	Excellent (N = 34)	Average (N = 71)	Poor (N = 34)	$F_{2,136}$ Ratio	P
Emotional Response to Height	10.90	12.10	12.70	4.61	.01
Unpleasant Experience	4.91	4.80	4.66	.72	.50
Pleasant Experience	3.85	3.81	3.68	2.15	.12
Expressed Fear of Height	2.30	2.47	2.44	1.22	.30
Risk-Taking Activity Level	10.85	8.96	8.19	8.45	.001
Running Skill	1.25	1.47	1.67	22.12	.001
Cross-stepping Skill	2.82	4.85	5.64	53.01	.001
Group Multivariate Means Axis I	2.30	2.77	3.26		
Axis II	6.06	6.32	6.06		

TABLE 3

MEANS, UNIVARIATE F RATIOS, AND PROBABILITY LEVELS FOR GROUPS CLASSIFIED BY PERFORMANC
ON HIGH CROSS-STEP TASK

Descriptive variables	Excellent (N = 35)	Average (N = 67)	Poor (N = 37)	$F_{2,136}$ Ratio	P
Emotional Response to Height	10.39	11.91	13.63	15.23	.001
Unpleasant Experience	4.82	4.81	4.72	.14	.86
Pleasant Experience	3.78	3.85	3.67	2.58	.08
Expressed Fear	2.28	2.46	2.48	1.57	.21
Risk-taking Activity Level	10.11	8.95	8.13	8.82	.00
Low Running Skill	1.23	1.45	1.73	33.90	.0
Low Grapevine Cross-stepping Skill	3.49	4.43	5.31	13.80	.0
Group Multivariate Means Axis I	2.30	2.77	3.26		
Axis II	6.06	6.32	6.06		

Cross-step Task

When the Ss were categorized into excellent, average, and poor performers on the basis of their scores on the more complex high cross-step task, again only one discriminant function was significant (see Table 1). This group of variables accounted for 98 percent of the variance and included only the Ss' current risk-taking activity level and their running and cross-step skill. Subjects' past experience with heights, and their expressed fear of height combined to account for only 2 percent of the total variance. Significant F values, displayed in Table 3, indicate group differences consistent with these findings. Although emotional response to height did not load high enough on the discriminant function to be considered a part of that cluster, the three group means were significantly different on that variable.

DISCUSSION

Subjects who were grouped on the basis of their dynamic gross motor performance 4 feet above the floor were differentiated by the descriptive variables, current risk-taking activity level and motor skill in the tasks, operating together. In the simpler running task, emotional responses also differentiated the groups. These findings are compatible with Wherry's model that predicts complex skill to be more affected by high levels of APTS. In this study, the more complex task of cross-stepping at heights was best differentiated by the Ss' skill at cross-stepping rather than by emotional responses. The discriminant weights of skill that are shown in Table 1 indicate that its contribution to the discriminating power of the variables may be seen to play the most predominant role in the more complex task. In other words, emotional responses and life style may be significant in predicting group performance under stress, but as the task becomes more complex, Ss' actual skill in the task assumes a more and more important role.

The assumption that providing individuals with pleasant task-height experiences in their childhood will result in a more enthusiastic and successful performance in college physical education risk-taking activities appears unwarranted. There is, however, support for the concept that persons who have moderate levels of skill in an area will perform in unfamiliar and stressful situations requiring that particular skill better than an individual who is unskillful. Individuals who are unskillful are apt to generate unusually high levels of APTS, which in turn cause performance to deteriorate more.

The findings of this study indicate that the educator can best predict a student's behavior under the stress of height by knowing the student's skill and activity interests, rather than by knowing his background or by attending to his fears regarding the task. It further suggests that it is probably not the occurrence of isolated experiences, such as building tree houses and swinging on tree ropes into the swimming hole, that are important in determining the student behavior. Rather, it is what the child who participates in these activities becomes as a result of them. It substantiates the fact that every individual should have the opportunity to develop motor skills to such an extent that he perceives himself to be capable of any physical demands which may be placed upon him, whether these demands occur within the framework of a stressful situation or not.

REFERENCES

1. Cooley, W. H. ; Lohnes, P. R., Multivariate Procedures for the Behavioral Sciences, John Wiley and Sons, New York, 1962.

2. Drinkwater, B. L. ; Flint, M. M., "Response Speed and Accuracy During Anticipatory Stress," Journal of Motor Behavior, 1:220-232, 1969.

3. Martin, B., "The Assessment of Anxiety by Physiological Behavioral Measures," The Psychology Bulletin, 58:234-255, 1961.

4. Wherry, R. J., "Model for the Study of Psychological Stress," Aerospace Medicine, 87:495-499: 1966.

5. Wherry, R. J. Jr.; Curran, P. M., "A Model for the Study of Some Determiners of Psychological Stress: Initial Experimental Research," Organizational Behavior and Human Performance. 1:226-251, 1966.

6. Wyrick, W., "Effects of Task Height and Practice on Static Balance Performance," Research Quarterly, 40: 215-221, 1969.

7. Wyrick, W., "Relationship of Ankle Strength and Test Order to Static Balance Performance," Research Quarterly, 40: 619-624, 1969.

From DERS

THE HUMANIST TRADITION IN MODERN INDIAN EDUCATIONAL THOUGHT

By K. G. Sayaidain. With a foreword by John Guy Fowlkes. 1967

256 pages, $5.95

Presents some of the significant contributions made to educational thought by Tagore, Gandhi, Iqbal, Azad, Radhakrishnan, and Husain. Gives Western leaders an insight into Indian educational trends.

"Here we have a very able Indian scholar writing about Indian leaders as an Indian sees them functioning in India and on the international scene."—Prof. CLIFFORD S. LIDDLE, Korea.

STATEMENT OF OWNERSHIP, MANAGEMENT AND CIRCULATION
(Act of October 23, 1962; Section 4369, Title 39, United States Code)

Publisher: File two copies of this form with your postmaster.
Postmaster: Complete verification on page 2

Form Approved,
Budget Bureau No. 46-R029

1. DATE OF FILING	2. TITLE OF PUBLICATION
October 23, 1970	JOURNAL OF EXPERIMENTAL EDUCATION

3. FREQUENCY OF ISSUE
QUARTERLY

4. LOCATION OF KNOWN OFFICE OF PUBLICATION *(Street, city, county, state, ZIP code)*
@L) 2101 Sherman Avenue - Madison, Wisconsin 53701

5. LOCATION OF THE HEADQUARTERS OR GENERAL BUSINESS OFFICES OF THE PUBLISHERS *(Not printers)*
2101 Sherman Avenue - Madison, Wisconsin 53701

6. NAMES AND ADDRESSES OF PUBLISHER, EDITOR, AND MANAGING EDITOR

PUBLISHER *(Name and address)*
Wilson B. Thiede - 4825 Bayfield Terrace - Madison, Wisconsin

EDITOR *(Name and address)*
John Schmid - University of Northern Colorado - 1 Greeley, Colorado

MANAGING EDITOR *(Name and address)*
Sandra L. Lewis - 1505 Neveda Road - Madison, Wisconsin

7. OWNER *(If owned by a corporation, its name and address must be stated and also immediately thereunder the names and addresses of stockholders owning or holding 1 percent or more of total amount of stock. If not owned by a corporation, the names and addresses of the individual owners must be given. If owned by a partnership or other unincorporated firm, its name and address, as well as that of each individual must be given.)* DEMBAR EDUCATIONAL RESEARCH SERVICES, INC.

NAME	ADDRESS
	2101 Sherman Avenue Madison, Wisconsin
Walter A. Frautschi	29 Fuller Drive - Madison, Wisconsin
Dorothy J. Frautschi	29 Fuller Drive - Madison, Wisconsin
John J. Frautschi	221 Lakewood Blvd. - Madison, Wisconsin
Walter J. Frautschi	3206 Lake Mendota Drive - Madison, Wisconsin

8. KNOWN BONDHOLDERS, MORTGAGEES, AND OTHER SECURITY HOLDERS OWNING OR HOLDING 1 PERCENT OR MORE OF TOTAL AMOUNT OF BONDS, MORTGAGES OR OTHER SECURITIES *(If there are none, so state)*

NAME	ADDRESS
NONE	

9. FOR COMPLETION BY NONPROFIT ORGANIZATIONS AUTHORIZED TO MAIL AT SPECIAL RATES *(Section 132.122, Postal Manual)*
(Check one)

The purpose, function, and nonprofit status of this organization and the exempt status for Federal income tax purposes
☐ Have not changed during preceding 12 months
☐ Have changed during preceding 12 months
(If changed, publisher must submit explanation of change with this statement.)

10. EXTENT AND NATURE OF CIRCULATION	AVERAGE NO. COPIES EACH ISSUE DURING PRECEDING 12 MONTHS	ACTUAL NUMBER OF COPIES OF SINGLE ISSUE PUBLISHED NEAREST TO FILING DATE
A. TOTAL NO COPIES PRINTED *(Net Press Run)*	2,612	2,700
B. PAID CIRCULATION 1. SALES THROUGH DEALERS AND CARRIERS, STREET VENDORS AND COUNTER SALES		
2. MAIL SUBSCRIPTIONS	1,969	2,075
C. TOTAL PAID CIRCULATION	1,969	2,075
D. FREE DISTRIBUTION *(including samples)* BY MAIL, CARRIER OR OTHER MEANS	113	71
E. TOTAL DISTRIBUTION *(Sum of C and D)*	2,082	2,146
F. OFFICE USE, LEFT-OVER, UNACCOUNTED, SPOILED AFTER PRINTING	530	554
G. TOTAL *(Sum of E & F—should equal net press run shown in A)*	2,612	2,700

I certify that the statements made by me above are correct and complete.

(Signature of editor, publisher, business manager, or owner)
Arnold A. Caucutt
-Business Manager

POD Form **3526** July 1970

DIRECTIONS FOR J.E.E. CONTRIBUTORS

The Journal of Experimental Education publishes specialized or technical education studies, treatises about the mathematics or methodology of behavioral research, and monographs of major current research interest.

ABSTRACT REQUIRED

An abstract of not more than 120 words must accompany each manuscript. It should precede the text, be designated *ABSTRACT,* and conform to the following standards:

1. In a research paper, include statements of (a) the problem, (b) the method, (c) the data, and (d) the conclusions.

2. In a review or discussion article, state the topics covered and the central thesis.

3. Standard abbreviations may be used freely if the meaning is clear. Use complete sentences and do not repeat information which is in the title.

TEXT

Usually research articles should have a clearly organized order of presentation. The following elements and their order is a guide and is not intended to be prescriptive.

The Problem. The nature, scope, and significance of the problem should be presented.

Related Research. Presentation and discussion of related research should be selective and should include references essential to clarify the problem under examination.

Methodology. This section should consist of hypotheses, description of the sample and sampling procedures, discussion of the variables, description of the data-gathering instruments (if well-known, their description may be omitted), and general description of the statistical procedures.

Presentation and Analysis of Data. Analysis of the data and conclusions about the hypotheses should be more than mere presentation. Rival hypotheses and significance for related studies and educational theory should be considered. Summary data (means, standard deviations, frequencies, correlation coefficients, etc.) should be presented in tabular or graphic form. Discussion of these data should be interpretive.

Summarizing Statements. A summary of conclusions and implications for education may supplement the abstract.

STYLE

Certain suggestions are offered here to achieve uniformity in publication style and to conserve the time and energy of the author and editorial staff in the preparation of copy. *A Manual on Writing Research,* 1962, and *A Manual of Form for Theses and Term Reports,* 1962, by Kathleen Dugdale, the Indiana University Bookstore, Bloomington, may be used as style manuals in preparation of manuscripts.

Two Copies Required. Copies should be double spaced with wide margins. Typed copies are preferred, but dittoed or mimeographed copies will be accepted if they are legible.

Subheads. Articles are frequently improved by the judicious use of subheads. However, avoid use of the title, INTRODUCTION, for a lead section.

Title. Try to use a short title, preferably no more than ten words. Avoid superfluous phrases, such as "A Comparison of . . .," "A Study of . . .," and "The Effectiveness of . . . ."

Tables. Prepare tables precisely as they are to appear in *The Journal,* caption each with a brief and to-the-point title, and number consecutively with arabic numerals: *Table 3.* INTERCORRELATION MATRIX, VARIABLES OPTIMALLY ORDERED.

Figures. Draw any graphs and charts in black ink on good quality paper, title each, and number consecutively, thus: *Figure 4.* SCHOOL ENROLLMENT. Send the original copy. We cannot accept mimeographed figures or charts. Data should not be framed, and ordinates and abscissas should be shortened to occupy as little space as possible.

Tables and Figures. Tables and figures must be original copies acceptable for reproduction. A charge will be assessed for any redrawing or re-typing of tables or figures.

Technical Symbols. All symbols, equations, and formulas must be clearly represented. Differentiate between numbers and letters (zero and the letter o; one and the letter l, etc.) Identify hand-drawn Greek letters in the margin. Align subscripts carefully.

Footnotes. Avoid explanatory footnotes by incorporating their content in the text. However, for essential footnotes, identify them with consecutive superscripts, as *book,*[2] *study,*[3] etc., and list the footnotes in a section, entitled FOOTNOTES, at the end of the text, but preceding the REFERENCES.

References. References should be listed alphabetically according to the author's last name at the end of the manuscript. Each entry should be numbered. Citation of a reference in the text should be made with this number, as (2); or if specific pages are to be indicated, as (2:245–7).

Illustrations of article and book references:

1. Bittner, Reign H. and Wilder, Carlton E., "Expectancy Tables: A Method of Interpreting Correlation Coefficients," *The Journal of Experimental Education,* 14:245–52, March 1946.

2. Thomson, Godfrey, H., *The Factorial Analysis of Human Ability,* Houghton Mifflin Co., Boston, 1950, 383 pp.

COSTS

The publisher charges a contributor's fee of $6 per printed page of approximately 1,200 words, billed upon publication. Authors are charged for changes in tables, figures, or copy made when article is in camera-ready form. Each contributor will receive 10 complimentary copies of the issue in which his article appears. Reprints are charged at cost, and a price schedule will be sent to each contributor.

PROOFREADING

We will send you proofs for correction (with instructions for handling). Any major changes made in the proofs that were not incorporated in your original copy will be an added expense to you. (Errors that we make, naturally, will be at our expense.)

Keep an exact carbon copy of your manuscript so that if a question arises the editors can refer you to specific pages, paragraphs, or lines for clarification by letter.

SEND MANUSCRIPTS TO

John Schmid, Department of Research Statistics and Methodology, Colorado State College, Greeley, Colorado 80631.

Environmental Education

What It Does

Environmental Education lends focus and thrust to the emerging American search for an environmental ethic—a developing discipline concerned with elucidating all the relationships of humanity to the total environment. To help professionalize this new field, this Journal will regularly publish reports about how better to communicate appropriate ecological and economic facts and esthetic dimensions that will lay a basis for citizen action by clarifying public choices in land and water use, relating them to general values and social objectives, instilling a desire for constructive change, and suggesting practical guidelines based on integrated approaches.

What This Journal Is

Environmental Education is a new quarterly devoted to discovery and dissemination in the emerging field of multidisciplinary conservation communications. Through a recording of insight and information about pioneer research and interpretation projects and programs, this Journal will act as a catalyst in encouraging and exchanging effective studies and developments toward broad public ecological awareness, understanding, and action.

How

Environmental Education publishes primarily research articles, news items, project reports, book reviews, and critical essays designed to advance the scientific study of conservation communications and improve field practice in environmental education.

ENTERING ITS SECOND YEAR — NOW 48 PAGES AN ISSUE

DEMBAR EDUCATIONAL RESEARCH SERVICES, INC., BOX 1605, MADISON, WISCONSIN 53701

Inscribe Your Name Today

———— I want to be a sponsor. Please enter my subscription to *Environmental Education* for two (2) years at $14.

———— I'd like to try *Environmental Education* for a year. Please enter my subscription at $7.50.

———— I'm interested in *Environmental Education*. Please send me a copy of the current issue for $2.

———— I am a student. Please send me *Environmental Education* for a year at $5.

———— I enclose a manuscript for your consideration.

———— I enclose a check. ———— Bill me. Bill _____

NAME _____

ADDRESS _____

CITY _____ STATE _____ ZIP _____

I think the following friend(s) of mine would be interested in hearing about *Environmental Education*:

THE *Journal* OF
Experimental
Education

CONTENTS

Relationship of Teaching Methodology to Student Achievement

Arvil S. Barr, Founder

EDITOR AND PUBLISHER ● 1932-1962

Dr. Herbert S. Conrad

*I*T IS WITH deep regret that the Executive Editors of The Journal of Experimental Education announce the death of Dr. Herbert S. Conrad on May 4, 1970.

He was Senior Research Adviser in the Office of the Associate Commissioner for Research, Bureau of Research, U.S. Office of Education, and Chairman of the Internal Clearance Committee in the Bureau of Research.

Dr. Conrad was born January 7, 1904. He obtained his bachelor's degree (cum laude) at Columbia University in 1926, and his master's and doctorate in psychology at the University of California, Berkeley, in 1930 and 1931, respectively. His principal positions included Research Associate in the Institute of Child Welfare (now the Institute of Human Development), Assistant Professor of Education, and Associate Professor of Psychology, all at the University of California (1928-43); Research Psychologist and Technical Consultant, College Entrance Examination Board and Educational Testing Service, Princeton, New Jersey (1943-48). Most of this work was for the Bureau of Naval Personnel and was financed by the National Defense Research Committee. Since 1948, he served in various positions in the U.S. Office of Education: first, as Director of the Research and Statistical Service (later termed the Educational Statistics Branch), 1948-60, and concurrently Acting Assistant Commissioner for Research, October 1956 - December 1957; then, as Coordinator of Research, Division of Higher Education, 1960-62; and, successively, as Program Development Officer, Program Evaluation Officer, and Senior Research Adviser in the Bureau of Research, 1963 until his death.

Among other activities, Dr. Conrad served for 10 years as Editor of the Psychological Monographs (1948-58); as President of four Divisions of the American Psychological Association (Maturity and Old Age, 1949; Psychologists in Public Service, 1952-53 and 1962; Evaluation and Measurement, 1956; Educational Psychology, 1961); and as Vice-President of the American Association for the Advancement of Science (Chairman, Section Q—Education), 1964. He served as a member of the editorial board of Psychometrika since 1944, and he was a consulting editor of The Journal of Experimental Education since 1935. He was the author of numerous articles and monographs in the fields of child psychology, statistics, and educational measurement. Recently he edited a memorial volume in honor of Dr. Harold E. Jones, long-time Director of the Institute of Human Development at the University of California. The book was entitled Studies in Human Development (Appleton-Century-Crofts, 1966).

In May 1968, he received the Superior Service Award for "broadening and significantly improving the Office of Education's basic research and statistical programs, in supporting its legislative programs, and in advancing extramural research relations."

U.S. Commissioner of Education, James E. Allen says, "Dr. Conrad was held in the highest esteem by everyone in the OE and by his friends and colleagues throughout the field of educational research. Education has lost a devoted educator."

Arvil S. Barr, Founder

EDITOR AND PUBLISHER ● 1932-1962

Dr. Herbert S. Conrad

ℐT IS WITH deep regret that the Executive Editors of The Journal of Experimental Education announce the death of Dr. Herbert S. Conrad on May 4, 1970.

He was Senior Research Adviser in the Office of the Associate Commissioner for Research, Bureau of Research, U. S. Office of Education, and Chairman of the Internal Clearance Committee in the Bureau of Research.

Dr. Conrad was born January 7, 1904. He obtained his bachelor's degree (cum laude) at Columbia University in 1926, and his master's and doctorate in psychology at the University of California, Berkeley, in 1930 and 1931, respectively. His principal positions included Research Associate in the Institute of Child Welfare (now the Institute of Human Development), Assistant Professor of Education, and Associate Professor of Psychology, all at the University of California (1928-43); Research Psychologist and Technical Consultant, College Entrance Examination Board and Educational Testing Service, Princeton, New Jersey (1943-48). Most of this work was for the Bureau of Naval Personnel and was financed by the National Defense Research Committee. Since 1948, he served in various positions in the U. S. Office of Education: first, as Director of the Research and Statistical Service (later termed the Educational Statistics Branch), 1948-60, and concurrently Acting Assistant Commissioner for Research, October 1956 - December 1957; then, as Coordinator of Research, Division of Higher Education, 1960-62; and, successively, as Program Development Officer, Program Evaluation Officer, and Senior Research Adviser in the Bureau of Research, 1963 until his death.

Among other activities, Dr. Conrad served for 10 years as Editor of the Psychological Monographs (1948-58); as President of four Divisions of the American Psychological Association (Maturity and Old Age, 1949; Psychologists in Public Service, 1952-53 and 1962; Evaluation and Measurement, 1956; Educational Psychology, 1961); and as Vice-President of the American Association for the Advancement of Science (Chairman, Section Q—Education), 1964. He served as a member of the editorial board of Psychometrika since 1944, and he was a consulting editor of The Journal of Experimental Education since 1935. He was the author of numerous articles and monographs in the fields of child psychology, statistics, and educational measurement. Recently he edited a memorial volume in honor of Dr. Harold E. Jones, long-time Director of the Institute of Human Development at the University of California. The book was entitled Studies in Human Development (Appleton-Century-Crofts, 1966).

In May 1968, he received the Superior Service Award for "broadening and significantly improving the Office of Education's basic research and statistical programs, in supporting its legislative programs, and in advancing extramural research relations."

U. S. Commissioner of Education, James E. Allen says, "Dr. Conrad was held in the highest esteem by everyone in the OE and by his friends and colleagues throughout the field of educational research. Education has lost a devoted educator."

AN EDUCATIONAL PROCESS MODEL

FOR USE IN RESEARCH

JENNY R. ARMSTRONG
The University of Wisconsin

ABSTRACT

An analysis of the formal educational process utilizing a pseudo-systems analysis approach was undertaken to identify and define the major input components which account for or cause changes in the major output component of learning defined in terms of learner behavioral change. Approaches to the experimental study of this process as defined and delineated were suggested. The importance of studies which examine the higher order interaction areas was emphasized and justified on the basis of informational output maximization which accompanies experimental research of the higher order interaction areas of the input model.

CURRENT TRENDS toward increased commercial production of educational materials makes rational models for use in the analysis and thorough examination of the educational process extremely timely. Although research in education began some 73 years ago with the early survey studies of Rice (8), there has been a meager amount of consistent educational information and knowledge accumulated about the interaction of major input components and how these input components interact to affect pupil learning.

To a large extent, the failure of educational research to contribute large consistent bodies of knowledge about the educational process has been faulty experimental design (4). Another important factor contributing to the lack of consistent educational information has been the failure to realistically conceptualize a usable model which identifies and concisely defines the major input and output factors of the educational process which must be considered (either controlled or systematically varied) in the design of educational experiments. Thus, the following analysis of the educational process and a conceptualization of a basic model for use in the design of educational experiments was undertaken.

MAJOR INPUT ELEMENTS OF THE FORMAL EDUCATIONAL PROCESS

One of the factors which has directly contributed to the minimal informational output of educational experiments is the failure to consider all of the major input elements of the educational process. This is in part attributable to the failure of previously proposed theories to separate and specifically define some of the more pertinent factors which affect various types of learning. Six major components in the educational process uniquely and in combination have been shown to affect learning. These input components are: curriculum (C), instruction (I), teacher(s) and/or implementor(s) (T), learner(s) (L), media (M), and environment (E).

Traditionally, the tendency has been to combine or singly define in global terms curriculum and instruction. MacDonald (7) has suggested the impossibility of defending such a position, however, from a systems point of view. Both the curriculum component and the instruction component make unique and independent contributions in the input process resulting in different types of learning. Even so, few research studies have been designed which separate these two factors. The results of two experiments in mathematics education (1, 2) which did separate the curriculum and instruction factors, support the Macdonald (7) contention.

The curriculum component is uniquely defined as the content or subject matter information which is organized into communicable form to be conveyed to the learner in the form of an educational program. The curriculum, then, has such general characteristics as scope, structure, organization, and sequence. Scope refers to the exact content included in the program. Structure refers to the elements about which the content is organized

(e. g., areas of discipline, general topics, broad generalizations, or problems). Organization refers to the characteristics of ordering the structured content (e.g., spiraling—the revisiting of topics, areas, or ideas at increasing levels of sophistication at successive intervals throughout the program). Sequence refers to the order of the content in the program.

The instruction category of the model uniquely includes all different methods or approaches of implementing the curriculum. The discovery method and the expository method are examples of two of the more predominately used and supported instructional approaches. Another set of strongly supported instructional techniques are the behavioral modification strategies and methodologies. Although research on instructional methods has been vast, the majority of the studies done have confounded the instruction factor with the curriculum factor (1). The same curriculum, thus, can be implemented utilizing different methods of instruction.

The teacher and/or implementor is yet another input component in the educational process model. The teacher (implementor) category of the model uniquely includes all of the different implementors of the curriculum. The teacher-implementor being human takes on all of the human characteristics of personality, intelligence, aptitude, attitude, etc. The implementor, however, in our more modern age of technology, may be a machine. This may range from a slide projector to the more complicated computer-linked apparatus. In many cases research has used a machine as the implementor and attempted to generalize to teacher (human) implementors. There are obvious differences which should be taken into account in design of and generalization from research.

The learner category of the model uniquely includes all receptors involved in the formal educational process. The learners are characterized by all factors known about the human to be critical to the learning process. Therefore, possible learner variables would be intelligence, personality, aptitude, sex, general physical characteristics, and attitudes.

A fifth major input component, environment, has been suggested by the work of the School Environment Research Project (6). This dimension would include such variables as classroom social climate, physical climate (e. g., spatial arrangement of furniture, boards, apparatus, temperature of the room, humidity, lighting, extraneous noise levels, etc.), and size of the group interacting in the learning process. Seldom has this dimension been given enough credance in the design of educational experiments in classroom situations. Too often, one treatment has been implemented in one classroom and another in a different classroom. Differences in learning could either be attributed to differences in environmental factors between the two classrooms or between the two treatments. Little, if any, information about the educational process is gained from such poorly conceptualized experimentation.

The sixth and last major input component, media, has been suggested by the continual increase in the number of different types of instructional materials which are currently available for use in the classroom. The elements which would be included in this category would be such materials as textbooks, trade books, films, slides, filmstrips, manipulative aids (e. g., blocks, counters, science equipment), games, and charts. Media is not to be confused with curriculum. Curriculum is the content as it is organized, delimited with respect to scope, sequenced, and structured within the media (e. g., textbook, filmstrip, film, etc.).

MAJOR OUTPUT ELEMENT OF THE FORMAL EDUCATIONAL PROCESS

The major output element of the formal educational process is learning. The general criterion of the success of the combinational input variables is assumed, here, to be learning on the part of the learners or subjects involved in the study. This criterion is to be considered in its broadest sense and thus includes all of the possible types (e. g., cognitive, affective, and psychomotor) and levels of learning (3).

INTERACTION OF THE MAJOR INPUT ELEMENTS

In considering this model for its general use in the design of educational experiments, all of the possible input element combinations must be conceptualized (see Table 1 and Figure 1). The curriculum-learner (CL) input combination, for example, would include experiments designed to determine the curriculum characteristics which for certain types of learners would maximize a desired kind of and level of learning. The instruction-learner (IL) input combination would include experiments designed to determine the methods or approaches which for certain types of learners would maximize a desired kind of and level of learning. The curriculum-media (CM) input combination would include research designed to determine the types and kind of media (e. g., film, textbook, filmstrip, slides, etc.) which for various types of content would maximize a desired kind and level of learning.

The input combination (CILTME) would include the most definitive and information-producing experiments portrayed by the model. Certainly, as more information about individual components becomes available, research can more frequently be done on these higher order overlap areas.

This general model can also be done using research in different subject matter areas (see Figure 2). After the research is completed there is new information about each of the input components which can be used to improve and revise hypotheses about how the variables within each one of the input areas together and singly maximize different types and levels of learning.

APPLICATION OF THE MODEL TO THE DESIGN OF EXPERIMENTS

One of the first steps in designing any educational experiment, using this model, is to identify the area which will be studied (see Table 1 and Figure 1). The objectives of the study can then be stated in the form of questions to be answered, hypotheses to be tested or effects to be estimated. Some of the standard designs which have been proposed for use in the design of educational experiments are both experimental (5) and quasi-experimental (4). In utilizing any of these designs, however, the major factors
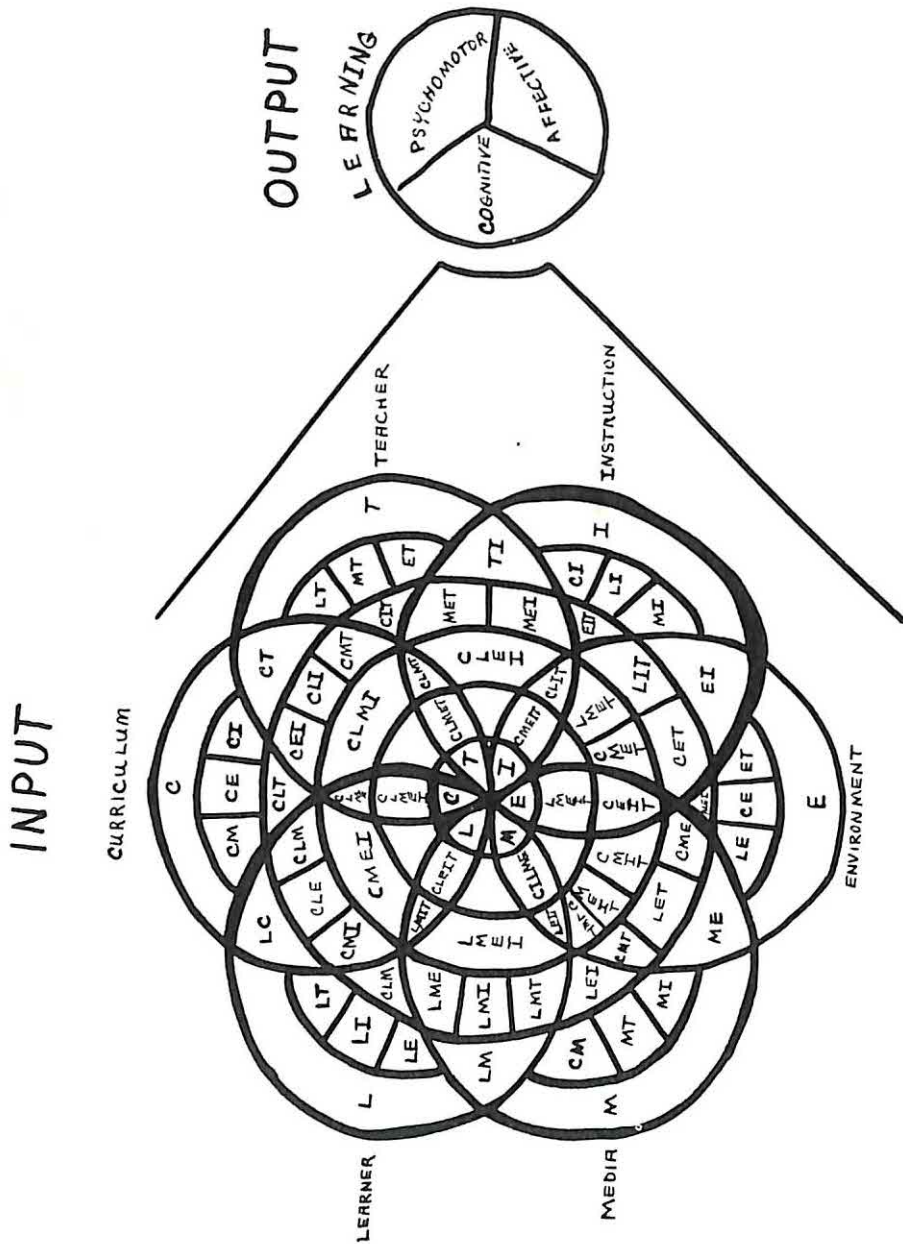
FIGURE 1

SCHEMATIC REPRESENTATION OF THE EDUCATIONAL PROCESS

TABLE 1

MAJOR INPUT COMPONENTS AND
COMBINATIONAL INTERACTIONS

Major Components:

Curriculum (C)
Instruction (I)
Teacher (T)
Learner (L)
Media (M)
Environment (E)

Logical Combinations:

Way:	Two	Three	Four	Five	Six
	CI	CIL	CILT	CILTM	CILTME
	CL	CIT	CILM	CILTE	
	CT	CIM	CILE	ILTME	
	CM	CIE	CITM	CLTME	
	CE	CLT	CITE	CILME	
	IL	CLM	CLTM	CITME	
	IT	CLE	CLTE		
	IM	CTM	CTME		
	IE	CTE	ILTM		
	LT	CME	ILTE		
	LM	ILT	ILME		
	LE	ILM	ITME		
	TM	ILE	LTME		
	TE	ITM	CIME		
	ME	ITE	CLME		
		IME			
		TLM			
		TLE			
		TME			
		LME			

outlined in this model should be taken into account so that confounding is minimized, so that meaningful information about the educational process can be derived from the ensuing research.

Another critical factor in the application of this general model is the generalizability and potential power of selected component variables. Generalizability, here, refers to the potential use of information derived in future educational and experimental problem situations. Power refers to the relative importance of the single variable in accounting for or affecting learning. If a variable accounts for only a small percentage of the learning (e.g., angle of the child to projection screen; taking a pretest over the subject matter to be taught), then certainly other more potent variables (e.g., sequential and/or hierarchical structuring of content; reinforcement structures and/or paradigms) should be the focus of the educator and/or experimentor's attention.

IDENTIFICATION AND INSTITUTION OF NEEDED CONTROLS

Once the area of research has been isolated, the next step is to identify the component areas which are to be controlled. By using the model, identification becomes simple. If the CIL area is to be studied, then the external control areas are M, E, and T. The internal control areas are C, L, and I. There are four ways to control variation: (1) by the general layout of the experiment, (2) the design of the experiment, (3) the statistical analysis of the experimental data, and (4) the development of the experimental program.

Consider this example. Suppose the variables of concern were learner cognitive development, curriculum sequence, and instructional approach, thus, placing the experiment in the CIL area of the model. Since the generalizability potential of the variables chosen for study is critical, systematic and logical analyses should be done regarding each of the overlap or interaction hypotheses of the variables chosen. Suppose, for example, the variables chosen in the CIL area were: curriculum sequence – psychological sequence paradigm versus subject matter sequence paradigm; instructional approach — discovery versus expository; and learner cognitive development — sensorimotor stage, stage of concrete operations, period of formal operations. In all cases of overlap or interaction, reasonable alternative hypotheses can be stated and supported. Consequently, the potential informational output of such an experiment is maximized. Furthermore, the general variables selected have maximal generalizability potential to the structuring of future learning environments.

Since cognitive development cannot be randomly assigned to learners, this must be considered a blocking variable. Learner(s) consequently must be grouped according to their manifest cognitive development level and then subjects within the three levels randomly assigned to the four treatment combinations. This yields a standard 3x2x2 randomized block design with several replicates per cell.

In order to internally control for other learner variables, subjects within each block would be, as the randomized block design (5) specifies, randomly assigned to each of the four treatment conditions (i.e., 1) subject matter sequence-discovery approach, (2) subject matter sequence-expository approach, (3) psychological sequence-discovery approach, (4) psychological sequence-expository approach). A sketch of the design is shown in Figure 3.

To control or minimize the effects of other instruction variables, the instructional programs must be designed so that the only difference between the two implementation procedures is the mode of presenting the curricula. Thus, either teachers must be trained
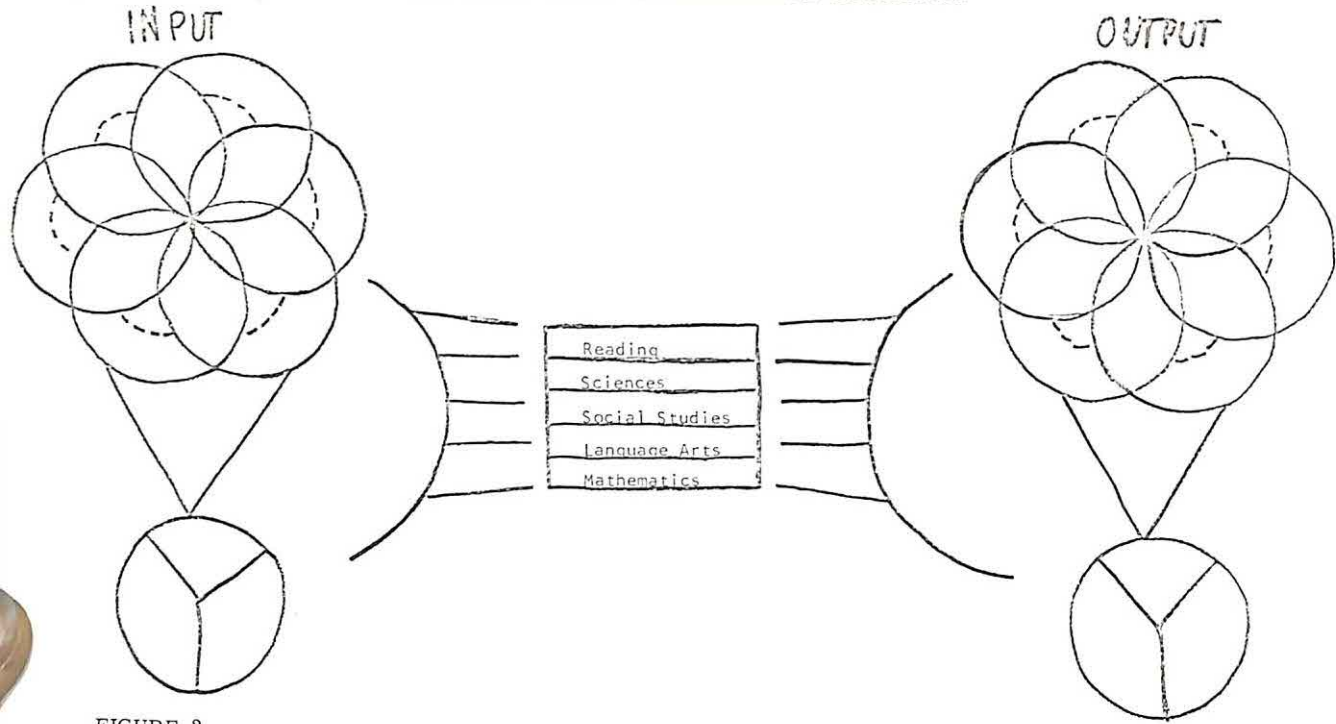
FIGURE 2

UTILIZATION OF THE RESEARCH MODEL IN DIFFERENT CONTENT AREAS
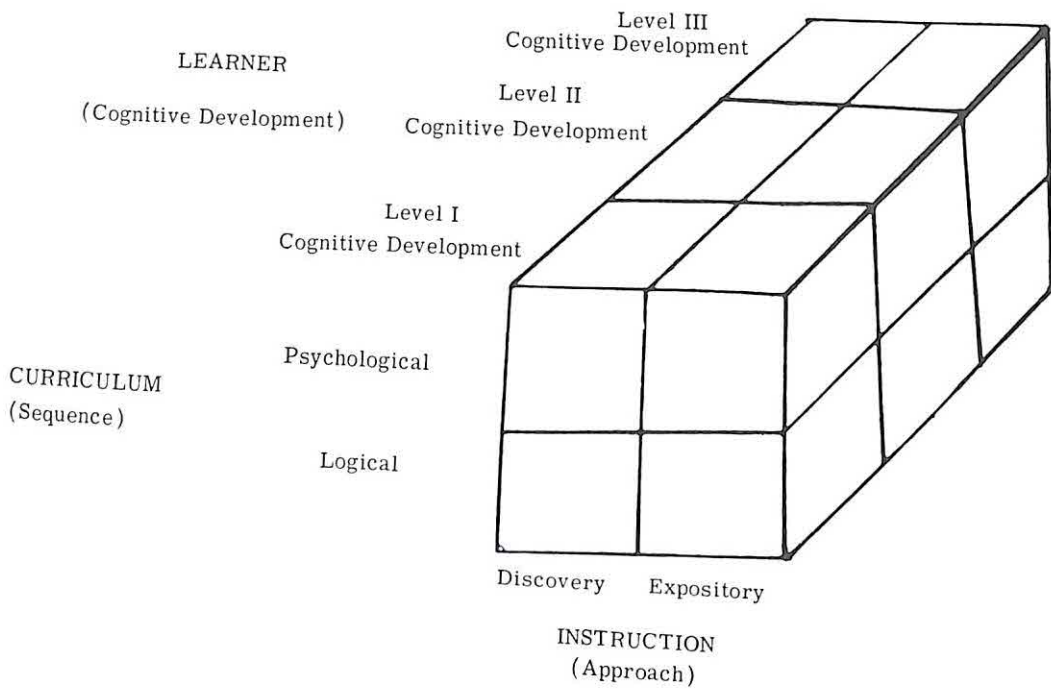
FIGURE 3

EXEMPLARY SKETCH OF "CIL" STUDY UTILIZING A 3x2x2 RANDOMIZED BLOCK DESIGN

to implement the curriculum using exactly the same techniques in all instances except for the mode of presentation or a machine of some type must be programmed such that the only differences are in mode of presentation. Generalization, naturally, is quite different in the two cases. Teachers as implementors are quite susceptible to error in experimental situations. Even so, the added flexibility that they bring to the teaching situation can be brought to bear on instances a priori unforeseeable when programming a machine to implement the program. Certainly, due consideration should be given to the advantages and disadvantages of these two alternatives.

The external components, those major components in the model not being specifically studied, T, M, and E, must also be considered. The effects of the T component can best be controlled or minimized by random assignment of (human) teachers to all treatment conditions if the number of teachers is large, or random assignment and systematic rotation of (human) teachers to the various treatment conditions if the number of teachers is small. Another alternative is to use a nonhuman implementor or machine. If a machine is used, the same type of machine must be used across all treatment conditions.

The M component can be controlled by equalizing the use of various types of media across the four treatment conditions. If manipulatives are used in one treatment condition, then they must be used at precisely the same time and for the same purpose in each of the other treatment conditions. To insure this type of control, a large amount of pre-experimental program planning must take place either via programming for machines or training of teachers.

The E component is much more difficult to control, but is causing increased concern on the part of educators. The effects of physical characteristics of the environment can usually be minimized by equalizing the classroom or instructional environment across treatment conditions.

If experiments are to be carried out in regular classrooms, the manipulation of environmental conditions for research purposes may be difficult. One solution, again, is the random assignment of experimental treatment conditions to classrooms and then systematic rotation of subjects within treatment conditions to all of the various classroom environments. Another possibility is the utilization of an experimental classroom on wheels. In this way, all treatments across schools and across classrooms would be in the same environment for all treatment conditions. In other words, the environment would remain constant across treatment conditions.

The effects of the social environment are not so easily minimized. Reduction to a large extent can be handled by randomly assigning subjects to treatment groups. One might assume, then, that social interaction patterns of various types would have an equally likely and mutually exclusive chance of occurrance in each of the various treatment groups.

The example given here is only one of innumerable studies which could be conducted using this general research model. Hopefully, as more information about the educational process is gained through research of this type, research in the higher order interaction areas can be completed.

To maximize the learning potential of the learner involved in the formal educational process, the major input element components must be interacting in an ideal fashion at a particular point in time. Ultimately, there may come a time when the educator knows how to amalgamate media, curriculum, instruction, environment, teacher(s) and/or implementor(s), and learners to maximize a desired kind and level of learning. The information needed can only come through systematic and controlled research of the formal educational process.

REFERENCES

1. Armstrong, Jenny R., "The Relative Effects of Two Forms of Spiral Organization and Two Modes of Presentation on Mathematical Learning," unpublished doctoral dissertation, University of Wisconsin, Madison, 1968.

2. Armstrong, Jenny R., "The Relationship of Mathematics Curriculum Innovation Presented Through Two Methods and Effect Upon Achievement as a Function of Learner Ability," Final Report, Part II, U.S. Office of Education, April 1968, unpublished.

3. Bloom, Benjamin S. (ed.), Taxonomy of Educational Objectives Handbook I: Cognitive Domain, McKay Co., Inc., New York, 1956.

4. Campbell, Donald T.; Stanley, Julian C., "Experimental and Quasi-experimental Designs for Research on Teaching," Handbook of Research on Teaching, 5:171-246, Rand McNally, Chicago, Illinois, 1963.

5. Cochran, William G.; Cox, Gertrude M., Experimental Designs, Second Edition, John Wiley and Sons, Inc., New York, 1957.

6. Larson, C. Theodore, Environmental Analysis, School Environments Research Publication Number 3, University of Michigan, Ann Arbor, 1965.

7. Macdonald, James B., "Educational Models for Instruction—Introduction," Theories of Instruction, Association for Supervision and Curriculum Development, Washington, D.C., 1965.

8. Rice, Joseph M., "The Futility of the Spelling Grind, I and II," Forum, 23:163-172; 409-419, 1897.

ON SCORING MULTIPLE CHOICE EXAMS

ALLOWING FOR PARTIAL KNOWLEDGE

J. C. Arnold and P. L. Arnold
Virginia Polytechnic Institute, Blacksburg, Virginia

ABSTRACT

A scoring procedure for multiple choice exams which allows for partial knowledge and also allows the examiner to control the expected gain due to guessing is considered in this paper. The procedure is considered from an elementary game theory approach. Comparisons are given with other scoring methods.

THE LITERATURE for mental test theory is a vast one and interest and research is accelerating. Procedures for scoring examinations, particularly multiple-choice tests, have received considerable attention in these research efforts. Among several others these include papers by Chernoff (3), Horst (6), Guilford (5), Calandra (2), and Arnold (1). In recent literature (see DeFinetti (4)), the personal probability approach has been considered for evaluating the knowledge of an examinee. However, the use of personal probabilities is a controversial area of scientific inference. Two basic problems to be considered in scoring procedures for multiple-choice tests are: (1) How does one handle the guessing factor? (2) How much of the available information has been recovered?

In this paper we consider a scoring procedure for multiple-choice exams which allows for partial knowledge and also allows the examiner to control the expected gain due to guessing. The procedure considered here is presented from an elementary game theory approach. The proposed method has been successfully used in actual classroom examinations in more than one subject matter. The results of one such classroom examination, using the proposed scoring procedure, are compared with several well known methods of scoring multiple-choice tests. Possible explanations for differences between the various methods are also given.

PROBLEM

We will formulate the problem in terms of a single item. The composite score for an examination consisting of several items will usually consist of adding the scores for each individual item. Suppose that a question (item) has k possible choices, one best (correct) choice and (k-1) distractors. The instructions to the examinee will be to give the smallest subset that he can give of which he is confident contains the correct response; i. e., the examinee is simply instructed to mark out as many incorrect responses as he can for each item. He should also be informed (for the purpose of discouraging outright guessing) that he receives a penalty if the correct or "best" response is marked out as an incorrect answer. This procedure then allows the examinee to receive credit for partial knowledge, the credit being greater the more incorrect responses than he can eliminate. It will readily be seen that such a procedure is simple to score. Simplicity in scoring is usually considered a desirable property when multiple-choice exams are utilized.

The primary concern here is to determine the appropriate scores that should be assigned to the various possible outcomes which may be given by an examinee, and also to correct for the guessing factor which is always an important consideration in multiple-choice tests. Arbitrarily assigning scores to the various outcomes would be undesirable. We will illustrate our approach to this problem by way of an example. Suppose that an item has four possible choices, one of which is correct and three of which are distractors. Suppose also that the examinee is instructed to delete as many incorrect answers as possible for each item. Looking at the problem as an

elementary game, let us consider the possible "strategies" available to the examinee. Suppose that the credits (scores) to be given to the examinee for each possible outcome are as follow:

C_0, if examinee eliminates no distractors,

C_1, if examinee eliminates one distractor,

C_2, if examinee eliminates two distractors,

C_3, if examinee eliminates all three distractors,

-P, if examinee eliminates the correct response as a distractor.

The reasonable strategies available to the examinee will depend upon his true state of knowledge relative to the number of distractors that he is able to eliminate. For example, if the examinee's true state of knowledge is such that he is unable to eliminate any of the distractors, then he has the choices of (a) simply not guessing, (b) attempting to eliminate one of the three distractors at random for which he has three chances out of four of being successful, (c) attempting to eliminate two of the three distractors at random for which his chances are two out of four of being successful, or (d) attempting to eliminate all three distractors at random for which his chances of being successful are reduced to one out of four.

The available strategies for each possible state of knowledge for an item with one correct answer and three distractors are listed at I, II, III, and IV, below.

I. If the state of knowledge is such that no distractors can be eliminated, the possible strategies are:

(i) do not guess and take a sure credit of C_0.

(ii) eliminate one choice at random for a possible credit of C_1 with probability 3/4 or a penalty of -P with probability 1/4.

(iii) eliminate two choices at random for a possible credit of C_2 with probability 1/2 or a penalty of -P with probability 1/2.

(iv) eliminate three choices at random for a possible credit of C_3 with probability 1/4 or a penalty of -P with probability 3/4.

II. If the state of knowledge is such that one distractor can be eliminated, the possible strategies are:

(i) do not guess and take a sure credit of C_1.

(ii) eliminate one choice at random for a possible credit of C_2 with probability 2/3 or a penalty of -P with probability 1/3.

(iii) eliminate two choices at random for a possible credit of C_3 with probability 1/3 or a penalty of -P with probability 2/3.

III. If the state of knowledge is such that two distractors can be eliminated, the possible strategies are:

(i) do not guess and take a sure credit of C_2.

(ii) eliminate one choice at random for a pos-

sible credit of C_3 with probability 1/2 or a penalty of -P with probability 1/2.

IV. If the state of knowledge is such that all three distractors can be eliminated, then full credit of C_3 is received with probability 1.

The possible strategies for items with other than four choices could be enumerated in a similar manner. For the above example of four choices, we will now calculate the correct scores C_0, C_1, C_2, and C_3 so that the game of guessing is a "fair game"; i.e., so that if the examinee wishes to guess and gamble for more credit at the risk of a penalty, then his expected gain due to guessing is 0. Procedures which may be used to penalize the guesser so that his expected gain due to guessing is negative (i.e., it does not pay to guess) will also be discussed.

We first note that expected gain due to guessing for a given strategy is:

EG = Expected credit when the game of guessing is played − Expected credit when the game of guessing is not played,

where the expected credit due to guessing is given by

[credit if win game] P (win game) - [penalty if lose game] P (lose game).

Hence, for a four-choice item, if the examinee cannot eliminate any of the distractors and we require a 0 expected gain due to guessing, then for the four strategies given at (I) the expected credit is, respectively

(i) $(C_0)(1) - (P)(0) = C_0$,

(ii) $(C_1)(3/4) - (P)(1/4) = 1/4(3C_1 - P)$,

(iii) $(C_2)(1/2) - (P)(1/2) = 1/2(C_2 - P)$,

(iv) $(C_3)(1/4) - (P)(3/4) = 1/4(C_3 - 3P)$,

where -P is the penalty for incorrectly marking the correct answer as a distractor. In order that the gain due to guessing be 0, we require that the expected credit for each of the above four cases be equal to C_0, which is the score given to the examinee who cannot eliminate any of the distractors and who decides not to guess. But for a 0 gain due to guessing we must therefore require $C_0 = 0$ since C_0 is the score given when the examinee cannot eliminate any of the three distractors. Hence, the "fair" score for four choices to an item is

$$C_0 = 0, \quad C_1 = P/3, \quad C_2 = P, \quad \text{and} \quad C_3 = 3P,$$

where P is the penalty for eliminating the correct answer as a distractor.

For the purpose of illustration, let us suppose that three points are to be assigned to C_3 (i.e., the credit received by an examinee who is able to eliminate all three distractors), then the penalty P should be -1 and the credit assigned to C_0, C_1, and C_2 should be 0, 1/3, and 1, respectively. Using these scores, we indicate below the appropriate credit to be given to an examinee for each of the fifteen possible outcomes for an item with four choices. An "x" beside a choice

indicates that the examinee has eliminated that choice as the correct answer. We suppose in the following example that the choice "c" is the correct answer to the item.

a. x	a. x	a. x	a.	a. x
b. x	b. x	b.	b. x	b.
c.	c.	c.	c.	c.
d. x	d.	d. x	d. x	d.

| 3 points | 1 point | 1 point | 1 point | 1/3 point |

a.	a.	a.	a.	a. x
b. x	b.	b.	b.	b.
c.	c.	c.	c. x	c. x
d.	d. x	d.	d.	d.

| 1/3 point | 1/3 point | 0 points | -1 point | -1 point |

a.	a.	a. x	a.	a. x
b. x	b.	b. x	b. x	b.
c. x	c. x	c. x	c. x	c. x
d.	d. x	d.	d. x	d. x

| -1 point | -1 point | -1 point | -1 point | -1 point |

From the above illustration it is easily seen that even though the number of possible outcomes are several, the examiner can determine the correct credit to be given very quickly by simply checking to see if the correct answer has been eliminated, and if not, by counting the number of distractors correctly eliminated.

We will now show that these scores are also the "fair" scores that should be given for each of the other possible states of knowledge considered above. For the three strategies given at (II), where the examinee is able to eliminate one distractor, the expected credits are, respectively

(i) $(C_1) (1) - (P) (0) = C_1$,

(ii) $(C_2) (2/3) - (P) (1/3) = P/3 = C_1$,

(iii) $(C_3) (1/3) - (P) (2/3) = P/3 = C_1$.

Hence, the expected gain due to guessing used in strategies (ii) and (iii) is 0 (i. e., the expected credit due to guessing is the same as that for correctly eliminating one distractor). Similarly if the state of knowledge is such that the examinee can eliminate two distractors (case III), then the expected credits for the two possible strategies are

(i) $(C_2) (1) - (P) (0) = C_2$,

(ii) $(C_3) (1/2) - (P) (1/2) = P = C_2$.

Of course, if the state of knowledge is such that the examinee can eliminate all three distractors, he receives the maximum credit of $C_3 = 3P$.

We will now consider the proper credit for the more general case. Suppose that the item of interest has k possible responses, (k-1) of them distractors and one correct or "best" answer. Let d denote the number of distractors eliminated by the examinee. We note that d can assume any of the values $0, 1, 2, \ldots,$ (k-1). If the true state of knowledge is such that the examinee cannot eliminate any of the (k-1) distrac-

tors, then he has the choice of randomly deleting d $(d = 0, 1, 2, \ldots, k-1)$ of them for a possible credit of C_d, and at the same time realizing the risk of receiving a penalty of $(-P)$ if the best answer is eliminated as a distractor. The examinee's expected credit for eliminating d responses at random is given by

$$(C_d) \left(\frac{k-d}{k}\right) - (P) \left(\frac{d}{k}\right), \qquad (1)$$

where

C_d = credit if d distractors are correctly eliminated,

$-P$ = penalty for eliminating the correct answer as a distractor,

k = total number of possible choices for item.

Equating the expected credit due to randomly guessing at (1) to 0, since 0 is the credit we would give to an examinee who is not able to eliminate any distractor and who also refuses to guess, and solving for C_d, we get

$$C_d = (P) \left(\frac{d}{k-d}\right). \qquad (2)$$

The credit C_d at (2) is the "fair" credit, in the sense of 0 expected gain due to guessing, for a multiple-choice question with k possible choices, (k-1) of them distractors, and -P the penalty of labeling the correct answer as a distractor.

If we consider the special case where partial credit is not allowed, then selecting only the "best" answer is equivalent to eliminating all (k-1) distractors. For this case (d = k-1), the credit is $C_{k-1} = (P) (k-1)$. Standardizing so that $C_{k-1} = 1$, we have $P = 1/(k-1)$ and the above scoring procedure simply reduces to the often used scoring rule of $S = R - [W/(k-1)]$, where R = number of correct answers, W = number of incorrect answers, and k = number of choices. This scoring rule is widely used in exams where the instructions are to encourage guessing only if at least one choice can correctly be eliminated. Hence, it is seen that this method is a special case of the procedure considered here where partial knowledge is allowed.

A summary of the "fair" scores for items with 2, 3, 4, or 5 choices is given in Table 1 with the standardized score (based on unity for full credit) written in parentheses.

We note in Table 1 that the case, where the number of choices per item is two, includes the true-false type of examination. Also, for convenience the scores given are standardized so that the maximum credit for each case is unity. Hence, if the examiner wishes to give a maximum credit greater than unity for correctly eliminating all of the distractors, he simply multiplies each of the possible credits for an item by the maximum credit given. For example, if an item has three choices and the maximum credit is to be 12 points, then C_0, C_1, C_2, and -P would become $0, 3, 12,$ and -6, respectively. Of course, the maximum credit to be given to an item would usually be the same as would be given for a correct answer if the usual scoring method of simply choosing only the correct answer was used.

TABLE 1

FAIR SCORES FOR VARIOUS STATES OF KNOWLEDGE

| | Number of Choices Per Item | | | |
	2	3	4	5
Fair Penalty	$-P$ (-1)	$-P$ $(-1/2)$	$-P$ $(-1/3)$	$-P$ $(-1/4)$
C_0	0 (0)	0 (0)	0 (0)	0 (0)
C_1	P (1)	P/2 (1/4)	P/3 (1/9)	P/4 (1/16)
C_2		2P (1)	P (1/3)	2P/3 (1/6)
C_3			3P (1)	3P/2 (3/8)
C_4				4P (1)

As stated, the scores given in Table 1 are the fair scores in the sense of 0 expected gain due to guessing. If the examiner prefers to discourage guessing, then increasing the "fair" penalty for specific scores causes the expected gain due to guessing to be negative; i.e., it does not pay the examinee to play the game, and hence, he should not guess. For example in the case of an item with four choices, the standard scores of 0, 1/9, 1/3, and 1 with a penalty of -1/3 is a fair game. When $P > 1/3$, the game of guessing will yield a negative expected gain due to guessing, the magnitude of which will depend upon how much larger than 1/3 the penalty P is specified. Hence, the examiner is able to control (or discourage) guessing to any degree desired by simply increasing the penalty for an incorrect answer. On the other hand, guessing could be encouraged by making the penalty smaller than the "fair" value.

COMPARISONS

The above scoring procedure was used for a classroom examination in elementary statistics. So that comparisons could be made with other scoring methods, the instructions were modified as follows:

For each question on this exam there are four choices for your answer, only one of which is correct. You are asked to mark out as many of the incorrect answers as you can. The more incorrect answers that you mark out the more credit you will receive. However, there is a penalty if you should mark out the correct answer. The credits and penalty for each question are calculated so that you have a 0 expected gain due to guessing. For each question, you are also asked to circle the one answer which you feel is most likely correct. Of course, if you are able to mark out all but one choice, then circle the only remaining answer.

The final instruction for circling the single choice which is considered most likely correct was needed in order that we could have a comparison with other

scoring methods. In our comparisons, we have treated the choice circled by the examinee for each question as the answer which would have been given if the instructions had been the usual ones for a multiple-choice test; i.e., to select only the correct answer for each question. The test consisted of nine questions, each with one correct answer and three distractors. The maximum credit was 9 points for 8 of the questions and 12 points for the other. From Table 1, we see that the "fair" scores are P = -3, $C_0 = 0$, $C_1 = 1$, $C_2 = 3$, $C_3 = 9$ for the 9-point questions and P = -4, $C_0 = 0$, $C_1 = 4/3$, $C_2 = 4$, $C_3 = 12$ for the 12-point question. Twenty-five students took the exam. Each question was attempted by every student.

The scoring methods used in our comparison with the partial knowledge procedure given above were:

(A) Full credit for each correct answer circled, with no penalty for guessing.

(B) Full credit for each question with correct answer circled but a penalty of -3 for the 9-point questions and -4 for the 12-point question for each item with an incorrect answer circled. Note that this is the conventional scoring formula given by

$$S = R - W/(k-1),$$

where R = number of correct responses, W = number of incorrect responses, and k = number of choices per question.

(C) $s = \lambda \left\{ \dfrac{kN_c - N}{N_c(k-1)} \right\}$, where $\lambda = 9$, or 12 for each

question, k = 4, N_c = number of correct responses to item from entire class, and N denotes the number of examinees attempting the question. We note that this score is the one which minimizes the mean square error about a true score of unity if the student actually knows the correct answer and 0 if the student does not know the correct answer (see 1).

TABLE 2

RELATIVE CLASS POSITION OF EXAMINEES FROM FIVE SCORING PROCEDURES

Examinee Number	Partial Knowledge	Scoring Method (A)	(B)	(C)	(D)
1	1.5	2	2	2	2
2	1.5	2	2	2	2
3	4	8	8	6.5	9.5
4	4	5	5	9	6
5	4	5	5	6.5	4.5
6	6.5	10.5	10.5	4	7.5
7	6.5	10.5	10.5	13	7.5
8	8	2	2	2	2
9	9	8	8	14	11
10	10	5	5	6.5	4.5
11	11	8	8	6.5	9.5
12	13.5	15	15.5	12	16
13	13.5	15	15.5	10	12.5
14	13.5	15	15.5	11	14.5
15	13.5	15	12	15.5	12.5
16	16	18	18	15.5	18
17	17	23	23	21.5	22.5
18	18	12	13	17	17
19	19	20	21	19	19.5
20	20	15	15.5	18	14.5
21	21	20	20	23.5	21
22	22	20	19	20	19.5
23	23	25	25	25	25
24	24	23	23	23.5	24
25	25	23	23	21.5	22.5

(D) Same as (C) except that the class was stratified on the basis of raw scores into the upper 25 percent, middle 50 percent, and lower 25 percent, and the score at (C) was calculated separately for each subgroup. This is recommendation (iii), of Arnold (1:11).

Each of the twenty-five students was then graded by the procedure given in this paper which allows for partial knowledge, and also by each of the methods at (A), (B), (C), and (D). The students were then ranked from high to low for each scoring method. In case of ties, average ranks were used. The ranks, representing relative scores for each of the five scoring methods given above are summarized for each student in Table 2.

It is of interest to note that as a general rule, the students who scored quite high with one scoring procedure also scored high with each of the others. Also, the ones who scored quite low with o n e method scored low with each of the other methods, too. However, the relative positions for the middle grades differ considerably from one scoring method to another. We feel that there are at least two reasonable explanations for this. Firstly, the random guessing factor causes the greatest variability i n this r a n g e. Secondly, since scores tend to cluster around the average the distance between ranks in the middle is usually less than those at either extreme. T h i s a l s o would cause greater variation for the ranks in t h e "middle" scores.

COMMENTS

We have considered a scoring procedure for multiple-choice tests where partial knowledge is allowed. The proper scores are well defined for any number of choices for an item so that there will be a 0 expected gain due to guessing. If it is desired to discourage guessing, one can simply enlarge the penalty beyond the "fair" value to any magnitude desired. If partial credit is not allowed this procedure reduces to the conventional score of $R - W/(k-1)$. The method has the practical advantage of being easily scored and the instructions are simple to understand. We also feel that this method enables more information to be extracted with multiple-choice exams. We have found the method favorably received by students taking exams using this scoring procedure.

FOOTNOTE

1. The authors wish to thank the Editor and a Con-sulting Editor for helpful comments which led to improvements in the presentation of this paper.

REFERENCES

1. Arnold, J.C., "On Scoring Procedures for Multiple Choice Tests," The Journal of Experimental Education, 38:9-12, 1969.

2. Calandra, A., "Scoring Formulas and Probability Considerations," Psychometrika, 6:1-9, 1941.

3. Chernoff, H., "The Scoring of Multiple Choice Questionnaires," Annals of Mathematical Statistics, 33:375-393, 1962.

4. DeFinetti, B., "Methods f o r Discriminating Levels of Partial Knowledge Concerning a Test Item," British Journal of Mathematical and Statistical Psychology, 18:87-123, 1965.

5. Guilford, J.P., "The Determination of Item Difficulty When Choice is a Factor," Psychometrika, 1:259-264, 1936.

6. Horst, P., "The Difficulty of a Multiple Choice Test Item," Journal of Educational Psychology, 24:229-232, 1933.

THE RELATIONSHIP OF WORK QUALITY IN UNDERGRADUATE MUSIC CURRICULA TO EFFECTIVENESS OF INSTRUMENTAL MUSIC TEACHING IN THE PUBLIC SCHOOLS

FRANCIS THOMAS BORKOWSKI
Ohio University

ABSTRACT

Subjects were high school instrumental music teachers. Grades and a self-rating served as measures of the quality of the Ss' undergraduate work. Thirty-one teaching effectiveness factors, consisting of pupil performance factors, pupil knowledge of music, band performance factors, and judgments by experts, were selected as an indication of teaching success. Each measure of undergraduate work was correlated to each teaching effectiveness factor and relations were inferred from the coefficients. It was found that quality of work in undergraduate courses is not related to teaching effectiveness as measured by pupil performance, pupil knowledge of music, or band performance. Quality of work in some undergraduate courses is related to experts' judgments of teaching effectiveness.

THE PURPOSE of this study was to determine the effect of work quality in undergraduate courses leading to a Music Education degree on effectiveness in teaching instrumental music to secondary school pupils.

The Ss were fifty-three West Virginia high school instrumental music teachers who were graduates of the four largest music schools in West Virginia. They had been teaching instrumental music in their present positions at least 3 years.

MEASURES OF UNDERGRADUATE CURRICULAR EXPERIENCES

Grades were used as a measure of the work quality by the S in undergraduate courses. Grades in specific courses were transcribed by direct examination of transcripts to sheets which categorized the separate courses into the following classifications: Music History, Music Theory, Major Performance Instrument, Minor Performance Instrument, Music Courses, Education Courses, Practice Teaching, Academic Courses, and Composite Average of all courses.

A self-rating on the S's level of performing ability was obtained and used as a measure of the quality of his undergraduate work. He was asked to rate himself on his major instrument performing ability, taking into consideration the highest position attained in a large musical organization, grades received on his major performance instrument, and difficulty of works that he performed. The self-ratings may have been biased by either vanity or humility; however, it can be assumed that the Ss have an intimate knowledge of how well they performed on their major performance instrument as undergraduates, and that this will be reflected in the self-ratings.

MEASURES OF TEACHING EFFECTIVENESS

In this study teaching effectiveness is determined by a description of learning in terms of specific performance abilities and levels of music knowledge which are considered indicative of the possession of that learning outcome. Thirty-one factors were selected as an indication of teaching effectiveness and consisted of learning outcomes of pupils. The teaching effectiveness factors were classified into the following four general categories: Pupil Performance Factors, Pupil Knowledge of Music History and Music Theory, Band Performance Factors, and Judgments by Experts.

Pupil Performance Factors

The experimenter visited the school in which each S was teaching and tape recorded eight pupils chosen by the S as representative of his best performers on each of the following instruments: flute, clarinet, saxophone, baritone, French horn, trumpet, trombone, and sousaphone. Each pupil was asked to play five musical examples that were specifically composed by the experimenter to determine the performance level of specific musical abilities that were later judged by three faculty members of the Division of Music at West Virginia University. Judges rated the quality of each pupil's performance on a 1-10 scale on the following factors: expressiveness of performance, rhythmic accuracy, fast staccato playing, fast legato playing, sight-reading, tone quality, and overall performance.

Pupil Knowledge of Music

After the pupils had performed the examples for the experimenter, they were given identical copies of a written Music Theory and Music History test. The Music Theory test comprised fifteen multiple-choice items. Correct answers required that the pupil know key signature, scales, triad formation, meters, and rhythms. The Music History test was divided into four sections. The first was concerned with the chronology of composers, the second with the identification of performers, the third with identification of major compositions, and the fourth with identification of major composers.

Band Performance Factors

The Ss were high school instrumental conductors. Each of their bands performed at either a regional or area music festival within a 2-week period during the course of the study. The experimenter tape recorded the S's band performance at one of the festivals. Three college instrumental music directors later judged the band performances on the following factors: tone quality, intonation, technique, balance, interpretation, and overall musical effect. Some of the several advantages in using the performance of these festival pieces as a measure of teaching effectiveness follow:

1. Each band is adequately rehearsed and at peak performance level.

2. Each band is sufficiently motivated by the possibility of receiving a superior rating.

3. Each band performs an adequate variety of styles from which aspects of performance can be judged.

Judgments by Experts

A composite rating was made of the overall teaching effectiveness of each S based on judgments by the following people who observed the S's quality of work: (1) the chairman of the music department from which the S received his undergraduate degree, (2) superintendent in the county where the S was then teaching, and (3) the experimenter. For the sake of convenience, the judges are referred to as experts. It is recognized that they are not all expert in each of the areas under consideration. They were, however, in a position to know the S's effectiveness as a teacher. They were asked to consider the degree of success of the S's performing groups and the abilities, attitudes, and knowledge which the pupils had derived from his instruction. It is noteworthy that there tends to be a high degree of agreement among the experts in regard to their judgment ratings in this study. It appears that common criteria are involved despite the difference of viewpoint and background.

Treatment of Data

There are nine measures of the S's quality of work in undergraduate courses and thirty-one teaching effectiveness factors. Each measure of undergraduate work was correlated to each teaching effectiveness factor while holding constant the S's years of teaching experience and the size of the school in which he was teaching. The data were analyzed by an IBM 1620 computer. The program yielded partial coefficients of correlation. In this study when a coefficient is at or above .265, thus the significance level is at or below .05, it is assumed that a relationship exists between the variables. In some cases in this study, when a number of correlation coefficients are individually nonsignificant in one classification but show a common trend, a binomial formula

$$\left(\tfrac{r}{n}\right)\left(\tfrac{1}{2}\right)r\left(\tfrac{1}{2}\right)n-r$$

was used to give the probability of this particular pattern of coefficients. If this probability is less than .05, the trend is adjudged as significant.

The relationship of the Ss' quality of undergraduate course work on their effectiveness in public school music teaching is inferred from the partial correlations and the levels of significance relating to those coefficients.

RESULTS

The results appear under the following headings representing the measures of teaching effectiveness used in this study.

Pupil Performance Factors

The coefficients of correlation (found in Tables 1 and 2) between the following measures of undergraduate courses and pupil performance factors are not appreciably higher than zero: music theory, major performance instrument, minor performance instruments, education courses, practice teaching, academic courses, music courses, composite average, and major performance instrument self-rating.

All of the correlations between music history grades and pupil performance factors resulted in negative coefficients. This particular trend is part of a pattern of negative coefficients when music history grades are used as a measure of undergraduate course work. The probability of this particular pattern of coefficients occurring is less than .01; therefore, the trend is adjudged as significant. Likewise, the trend of negative coefficients of correlation between education courses and pupil performance factors, and between a composite average of undergraduate grades and pupil performance factors, though individually nonsignificant, are part of a pattern adjudged as significant.

Pupil Knowledge of Music

The coefficients of correlation (found in Table 3) between the measures of undergraduate course work and pupil knowledge of music as measured by scores on music history and music theory tests are not appreciably higher than zero.

TABLE 1

PARTIAL CORRELATIONS BETWEEN UNDERGRADUATE GRADES AND PUPIL PERFORMANCE FACTORS ON ALL WIND INSTRUMENTS

	Expressiveness	Rhythmic Accuracy	Fast Staccato Playing	Fast Legato Playing	Sight-Reading	Tone Quality	Overall Performance (OP)	Mean Total Excluding OP	Mean Total Including OP
Music History Courses	-.326	-.248	-.304	-.182	-.216	-.216	-.249	-.286	-.263
Music Theory Courses	.067	.141	.061	.101	.079	.167	.096	.104	.099
Major Performance Instrument	.096	.066	-.039	.019	.020	.037	.013	.019	.008
Minor Performance Instrument	-.017	.052	-.012	.045	.040	.001	.032	.021	.024
Education Courses	-.224	-.091	-.260	-.097	-.052	-.208	-.140	-.152	-.149
Practice Teaching	.069	.062	.049	.113	.116	.079	.096	.077	.079
Academic Courses	-.047	-.047	-.075	.020	.048	.014	.028	-.026	-.016
Music Courses	.046	.005	-.127	.013	.000	.051	-.037	-.027	-.031
Composite Average	-.115	-.057	-.156	-.008	-.015	-.016	-.061	-.069	-.064
Major Performance Instrument Self-Rating	.183	.014	.113	.049	.103	.107	.149	.096	.100

The negative coefficients of correlation between music history grades and pupil music history or music theory scores, and between the composite average of undergraduate grades and pupil music history or music theory scores, though individually nonsignificant, are part of the pattern of negative coefficients adjudged as significant.

Band Performance Factors

The coefficients of correlation (found in Table 4) between undergraduate grades and band performance factors are not appreciably higher than zero. Six of eight coefficients that resulted from the correlations between the Ss' self-ratings on their major performance instruments and band performance factors are significant at the .05 level. The negative coefficients of correlation between music history grades and band performance factors, though individually nonsignificant, are part of the pattern of negative coefficients adjudged as significant.

Judgments by Experts

The largest coefficients of correlation (found in Table 5) in this study occurred between the following measures of undergraduate course work and judgments of the Ss' teaching effectiveness by experts: music theory grades, grades in music courses, and a composite average of undergraduate grades.

The coefficients of correlation between the following measures of undergraduate course work and judgments of the Ss' teaching effectiveness by experts are not appreciably higher than zero: music history grades, major performance instrument grades, and grades in education courses. It is noteworthy that the coefficient of correlation between music history grades and judgments by experts is not negative and thus deviates from the pattern of coefficients that results when music history grades are used as a measure of undergraduate course work.

Though individually nonsignificant at the .05 level, the coefficients of correlation between the following measures of undergraduate course work and judgments by experts are appreciably higher than zero: minor performance instrument grades, practice teaching grades, grades in academic courses, and a major performance instrument self-rating.

The coefficients of correlation between each measure of undergraduate course work and each classification of teaching effectiveness factors suggest that the criteria used by experts to judge teaching effectiveness are not based solely on the learning outcomes of the Ss' pupils. Though the experts may take into consideration these outcomes, factors apparently are operating on the teaching situation which influence their judgments, such as classroom discipline, personal attractiveness of the teacher, alertness of the pupils in the classroom, pupil morale, and possibly the economic environment. All of these foregoing factors and many others may be considered by the experts in the light of their own educational viewpoints.

TABLE 3

PARTIAL CORRELATIONS BETWEEN UNDER-
GRADUATE GRADES AND PUPIL KNOWLEDGE
OF MUSIC

	Pupil Music History Scores	Pupil Music Theory Scores
Music History Courses	-.016	-.153
Music Theory Courses	-.099	-.181
Major Performance Instrument	-.129	.059
Minor Performance Instrument	-.071	.016
Education Courses	.062	-.101
Practice Teaching	-.045	.119
Academic Courses	.072	.097
Music Courses	-.135	-.074
Composite Average	-.069	-.029
Major Performance Instrument Self-Rating	-.060	.016

TABLE 5

PARTIAL CORRELATIONS BETWEEN UNDER-
GRADUATE GRADES AND JUDGMENTS BY
EXPERTS

	Judgments by Experts
Music History Courses	.180
Music Theory Courses	.385
Major Performance Instrument	.145
Minor Performance Instrument	.248
Education Courses	.097
Practice Teaching	.251
Academic Courses	.231
Music Courses	.345
Composite Average	.381
Major Performance Instrument Self-Rating	.223

CONCLUSIONS

Quality of work in music history courses is not

TABLE 4

PARTIAL CORRELATIONS BETWEEN UNDERGRADUATE GRADES AND BAND PERFORMANCE FACTORS

	Tone Quality	Intonation	Technique	Balance	Interpretation	Overall Musical Effect (OME)	Mean Total of all Band Performance Factors	Mean Total Excluding (OME)
Music History Courses	-.153	-.214	-.121	-.134	-.191	-.155	-.162	-.160
Music Theory Courses	.156	.133	.157	.206	.198	.194	.183	.168
Major Performance Instrument	.019	.027	.033	.013	.126	.022	.047	.029
Minor Performance Instrument	.020	.055	.027	.074	.086	.066	.057	.058
Education Courses	-.010	-.006	.041	.033	.020	.010	.008	.006
Practice Teaching	.062	-.047	-.038	-.007	.029	-.008	.004	-.011
Academic Courses	.037	-.012	.056	.035	.062	.049	.041	.030
Music Courses	.128	.100	.115	.135	.156	.136	.134	.128
Composite Average	.090	.060	.111	.088	.129	.106	.102	.096
Major Performance Instrument Self-Rating	.334	.287	.248	.238	.304	.288	.294	.291

TABLE 6

PARTIAL CORRELATIONS BETWEEN UNDER-GRADUATE GRADES AND COMPOSITE SCORE OF PUPIL ACHIEVEMENT SCORES

	Composite Score
Music History Courses	-.201
Music Theory Courses	.002
Major Performance Instrument	-.006
Minor Performance Instrument	.020
Education Courses	-.058
Practice Teaching	.051
Academic Courses	.067
Music Courses	-.036
Composite Average	-.020
Major Performance Instrument Self-Rating	.121

related to teaching effectiveness as measured by judgments by experts. It is inversely related to pupil performance factors, pupil knowledge of music theory and music history, and band performance factors.

Quality of work in music theory courses, all music courses taken as an average, and all undergraduate courses leading to a Music Education degree taken as a composite average is related to teaching effectiveness as measured by judgments by experts. Quality of work in these courses is not related to pupil performance factors, pupil knowledge of music theory or music history, or band performance factors.

Quality of work in undergraduate courses, as measured by undergraduate grades and a major performance instrument self-rating, is not related to a composite score of pupil achievement scores, as evidenced by the size of the coefficients of correlation found in Table 6.

Quality of work on the Ss' major performance instrument, on the minor performance instruments, in practice teaching, and in the academic non-music courses is not related to teaching effectiveness as measured by the teaching effectiveness factors of this study.

Quality of work on the major performance instrument, as measured by a self-rating, is related to teaching effectiveness as measured by band performance factors. It is not related to judgments by experts, pupil performance factors, and pupil knowledge of music history or music theory. In comparing the results of the correlations between the major performance instrument self-rating and band performance factors with those that occurred between major performance instrument grades and band performance factors, it appears that the level of proficiency on the major performance instrument as viewed by the S may not be the same as the level of proficiency as viewed by his teacher.

EFFECT OF DETAILED GUIDANCE ON THE

WRITING EFFICIENCY OF COLLEGE FRESHMEN

PERRY R. CHILDERS and VIRGINIA J. HAAS
The University of Wisconsin-Milwaukee

ABSTRACT

This study was designed to determine the effectiveness of extensive pre-guidance, periodic checking, and detailed correction and comment on specific preliminary steps in the development of the documented research paper, written by freshman English students, according to their grade achievement. Also included was a cross-check analysis of written essays, prepared and impromptu. The study involved the research papers of forty-eight students enrolled in two classes. One group received detailed guidance, the other did not. An evaluation chart, developed by the instructor and another rater, was used on all papers; a systematic sampling of 19 percent of these was graded by the second rater. The inter-rater reliability coefficient (Pearson r) was established as .96.

The hypothesis—there will be no significant difference in the grade achievement of separate sections of second semester freshman English students on the research paper according to different instructional procedures—was not rejected. The conclusion was drawn that, within the limited scope of this study, the described methods were not validated by the classroom instructor as a means of improving the performance of students on the research paper.

THE EFFICACY of the college freshman English course at the University of Wisconsin-Milwaukee (UWM)—and at almost any university across the country, apparently (9)—is consistently questioned, often by the very people who teach the course. The practices questioned range from the rationale for impromptus to the reliability of the grading practices of individual instructors.

This investigation was prompted by concern about the practice of correcting and commenting on students' written work at length, especially the major documented paper required by the department in the second semester course. The question almost inevitably formed by the instructor was whether or not careful, precise marking of mechanical, structural, and formal errors on outlines and rough drafts, as well as lengthy end comment and individual conferences, would actually result in an improved final draft of the research paper. Obviously, careful, almost picayunish, attention to every possible error the instructor can find as well as one or two paragraphs of thoughtful comment on ideas and presentation of those ideas involves many hours of work for the instructor who is handling forty-five to fifty-five papers of perhaps eight to ten pages each. Equally obvious is the fact that it requires several hours of each student's time to produce the outline and rough draft and then to carefully note and attempt to refine a paper in terms of whatever comments seem valuable. In many cases, the process can also involve conferences between the instructor and student after submission of the outline and again after the rough draft is submitted.

It should be noted that the National Council of Teachers of English has pointed out that too little is known about composition teaching itself and that the body of research on composition is so slight and so primitive that it serves teachers of composition poorly (1). Not surprisingly then, this investigation discovered that extant research offered a body of information which clarified some portions of the question while suggesting conflict in other aspects of it. For example, the

Buxton (2) study indicates that college freshmen whose writing is thoroughly evaluated, criticized, and revised improve their writing more than their peers whose essays are not handled in this way. An experiment by Kincaid (10) showed significant variation in the quality of writing of freshmen ranked in the bottom quarter of his sample according to the theme assignment. And any number of studies, including those of Dye and Bledsoe (7), Cast (3), Fostvedt (8), Kincaid (10), Buxton (2), and Thompson (11), indicate that variations in grading practices significantly affect the mark a student receives on a piece of work or a semester's production; this mark, of course, must serve as a measure of the student's achievement in written work.

A review of studies which indicated that a multitude of variables can influence grade achievement by freshman English students and observations in freshman English sections suggested the following null hypothesis be tested:

> There will be no significant difference in the grade achievement of separate sections of second semester freshman English students on the documented research paper following different instructional procedures.

METHOD

The sample consisted of forty-eight students enrolled in second semester (English 102) English at UWM during the spring of 1969. They were enrolled in two sections. One section, Writing Group X, had twenty-five students; the other, Writing Group Y, twenty-three.

Kincaid's study (10) offers the precedent by which university registration procedures were relied on to insure that the sample was representative of this segment of the university population. Freshman registration is restricted to special days in the fall and spring term because of the massive numbers of freshman students. Each freshman registration form is stamped with a sequential number as it is received. Then the forms are collated according to preferred days and hours for freshman English, with the first received being placed in first section choices. Since the English department specifies a maximum of twenty-seven students in 102 sections, no more than this are assigned to sections beginning with the first section number for the specific period listed. Balance in number of students enrolled in a section is sought, but no attempt is made to distribute students according to any other standard. English 102 (which is required) includes those students who received at least a D in English 101. Those who receive an A in 101 are exempt from 102 but may take it if they wish. The preceding factors indicated the use of a systematic sampling (5) to obtain papers which were graded by a second rater.

Groups X and Y were given a schedule of deadlines for the research paper on the first day of class. All Ss were required to present a tentative topic during the third week of class. Students were asked to see the instructor for a conference or to submit a written statement if they preferred. The instructor and student either individually discussed narrowing of subjects, initial references, likely areas of difficulty and so forth, or the instructor returned similar suggestions in writing along with the S's declaration.

All Ss were asked to buy the English department pamphlet on the research paper and read it as well as two chapters on the research paper in the freshman rhetoric text. One 50-minute class period was devoted to an explanation of research paper form and purpose and to the technicalities of documentation. This was done on the same day for both groups.

Both groups handed in a short documented essay, the primary emphasis of which was the proper footnote and bibliography form, the incorporation of source material into the text, and documentation practices which would preclude any suggestion of plagiarism. Ss in Group X were advised, in marginal and end comment, of problem areas and were simply referred to pertinent pages in the pamphlet and rhetoric text. The papers of Ss in Group Y were given extensive marginal comment and suggestions, and those Ss with serious difficulties in coping with these procedures were called in for conferences.

During the fourth week of classes, Ss in Group X were informed that they would be required to submit nothing more than their final draft on the due date. They were also asked to turn in their note cards with their papers. Group Y continued to follow the original schedule, which required that note cards from at least three sources be turned in during this week. These cards were checked by the instructor to see if material was quoted properly, if the students were using fruitful sources, if the references pertained to the student's announced topic and thesis. Comments and questions were noted on the cards, and conferences were held at the request of students.

In the seventh week Group Y turned in sentence outlines of their proposed papers. These were read for content, organization, logic in the presentation of ideas, and for the inclusion of extraneous material. No grades were given. At this time the instructor requested conferences with specific students, of whom all but two responded.

At the end of the eighth week Group Y submitted rough drafts of their papers. These copies were read, by the instructor only, for all of the criteria that were checked in the final paper (see Figure 1). At least one paragraph—but sometimes two or more—of end comment indicated student strengths and weaknesses in rhetoric and form. Marginal comments were generally directed to pointing out specific instances of mechanical, formal, or structural error, although rhetorical considerations were not excluded from them. Individual conferences for some students in Group X were recommended. Conferences were also held with those members of the group who requested them.

During the twelfth week both groups submitted their papers. At the deadline, forty-five were turned in; three were submitted within the next three class periods.

In addition to the major documented paper, each group wrote seven essays. The assignments were the same for both groups. With the exception of the single theme already mentioned, none of the others required documentation or bibliography although some students, as a matter of choice wrote papers which incorporated source material. The essays were required to be the standard 500 words long. Students were given 7-10 days between time of assignment and due date of the theme except for two impromptu essays.

FIGURE 1

EVALUATION CHART (4)

RHETORIC

Control and Coherence: careful, complete development of individual's stated thesis; no extraneous material

Logic: understanding of content and presentation of it in an orderly way

Diction: appropriate word choices; effective level of usage; apt title

---

DOCUMENTATION

Reference Material: fairly wide selection of appropriate sources; pertinent selection of quotations and material

Incorporation: smooth integration of source material into student's text

Form: adherence to documentation requirements, footnote, and bibliography form as outlined in the department pamphlet

---

GRAMMAR AND STRUCTURE

Sentences: use of syntactically accurate, textured, appropriately varies ones

Paragraphs: use of orderly, unified, appropriately developed generative structures

Spelling, Punctuation, Transitions, Introduction, and Conclusion: use of techniques consistent with department standards and/or intent and meaning of writer

---

COMMENTS

TOTAL ALL SECTIONS

NUMERICAL GRADE

LETTER GRADE

NAME:

Reliability in grading the research paper was sought according to findings of prior experiments. Cast (3) and Buxton (2) suggested, respectively, that a weighted scale is most feasible for maintaining high reliability between raters, and that raters who develop the evaluative criteria together are likely to establish a high reliabilty rating. Additionally, the experiments of Fostvedt (8) and Cast (2), as well as the Work of Diederich (6), indicate the criteria most often relied on by raters of composition.

A second graduate teaching assistant, who came to UWM at the same time as the classroom instructor, was asked to cooperate in establishing an evaluation chart. The final jointly-developed chart was based on criteria selected from prior research, the particular nature of the assignment, and a standard departmental rating chart which is distributed to teaching assistants as a suggested guide.

Each of the three major divisions on this chart could receive a maximum of twelve points. Each subdivision was rated from 0-4, failing to superior. Both raters agreed that this system would allow any single student to excel enough in one area to achieve at least a passing grade on the paper and thus help to diminish the effects of any bias in the instructor's ratings, since the students could not be anonymous for the instructor as they were for the other rater.

The research papers were submitted in no particular order. The three later papers were inserted into the proper stack by an uninvolved individual. A systematic sampling of every fifth paper (approximately 19 percent of the total) was abstracted. The groups were not mixed, and the sampling consisted of four papers from one group and five from another. The volunteer rater did not know to which group the papers belonged. All the papers were then rated by the instructor. The inter-rater reliability coefficient was established as .96.

RESULTS

The null hypothesis was tested by means of the t-test, with the .05 level of significance set for rejection of that hypothesis.

The second analysis concerned differences which existed in the achievement of Groups X and Y on the regular essays.

Finally, using the t-test again, the achievement of Group X on the research paper and regular essays was compared; the same procedure was repeated for Group Y.

The t-value established for Group X and Group Y on the research paper was .7737 with 46 degrees of freedom—not significant. The null hypothesis was not rejected.

Further, the t-value for Groups X and Y on their regular essays was .9567, again not significant.

The within-group comparison yielded a t-value for Group X on the research paper and regular essays of .7398 with 48 degrees of freedom, not a significant difference. The value for Group Y was 1.1587—not significant.

The tables should be studied with this in mind: the first semester is entirely devoted to the single short essay. Ten essays are required during the first semester, and classroom discussion and outside reading are directed to techniques of producing acceptable short essays. It can also be noted that there is an almost exact reversal in the mean and standard deviation figures for Groups X and Y on the regular essays and the research papers (see Table 1).

CONCLUSIONS

Since the null hypothesis was not rejected, it was concluded that the described process of periodic su-

TABLE 1

COMPARISON OF ACHIEVEMENT ON THE RESEARCH PAPER AND REGULAR ESSAYS BETWEEN GROUPS

	Writing Group X			Writing Group Y				
	N	$\overline{X}$	s	N	$\overline{X}$	s	df	t-value
Research Paper	25	5.64	3.26	23	6.39	3.45	46	.773
Regular Essays	25	6.04	2.41	23	5.39	2.29	46	.957

pervision and extensive correction and comment was not validated as a means of aiding students to produce a more effective research paper than their peers, who have not received such supervision and comment.

The additional between-group test for a significant difference in achievement on the regular essays was carried out in an attempt to determine whether a difference in the type of assignments would affect either of the writing groups. Since this difference was not significant, it was concluded that a possible major variable was minimized. The comparison also indicated that neither group was a significantly superior writing group (see Table 2).

TABLE 2

COMPARISON OF ACHIEVEMENT ON THE RESEARCH PAPER AND REGULAR ESSAYS WITHIN GROUPS

	Writing Group X		Writing Group Y	
	Research Papers	Regular Essays	Research Papers	Regular Essays
N	25	25	23	23
$\overline{X}$	5.64	6.04	6.39	5.39
s	3.26	2.41	3.45	2.29
df	48		44	
t-value	.740		1.159	

DISCUSSION

A greater number of raters would have been desirable in this experiment because the nature of the procedures precluded anonymity for the student-writers as far as the instructor was concerned. However, the extremely high reliability rating coefficient indicates, for these samples at least, that a variation in grade achievement because of grading practices or bias was minimized. Peripherally, this high correspondence agrees with the findings of prior research which states that high reliabilty between raters can be achieved when raters select criteria together and are familiar with the application of chosen standards.

REFERENCES

1. Braddock, Richard; Lloyd-Jones, Richard; Schoer, Lowell, Research in Written Composition, National Council of Teachers of English, Champaign, Illinois, 1963, pp. 5, 65-93.

2. Buxton, Earl W., "An Experiment to Test the Effects of Writing Frequency and Guided Practice Upon Students' Skill in Written Expression," in Braddock; Lloyd-Jones; Schoer, Research in Written Composition, National Council of Teachers of English, Champaign, Illinois, 1963, pp. 65-81.

3. Cast, B. M. D., "The Efficiency of Different Methods of Marking English Compositions," British Journal of Educational Psychology, 9:257-269, 1939 and 10:49-60, 1940.

4. Coward, Ann F., "A Comparison of Two Methods of Grading English Compositions," The Journal of Educational Research, 46:81-93, 1952.

5. Croxton, F. E.; Cowden, D. J., Applied General Statistics, Prentice-Hall, Inc., Englewood Cliffs, New Jersey, 1955, pp. 27-28, 645-650, 750-751.

6. Diederich, Paul B., "How to Measure Growth in Writing Ability," English Journal, 55:435-449, 1966.

7. Dye, James M.; Bledsoe, Joseph C., "An Experiment with Grouping of Freshman English Composition Students," The Journal of Experimental Education, 35:71-74, 1967.

8. Fostvedt, Donald R., "Criteria for the Evaluation of High School English Composition," The Journal of Educational Research, 59:108-112, 1965.

9. Gerber, John C. (Ed.), The College Teaching of English, Appleton-Century, Meredith Press, New York, New York, 1965.

10. Kincaid, Gerald L., "Some Factors Affecting Variations in the Quality of Students' Writing," in Braddock; Lloyd-Jones; Schoer, Research in Written Composition, National Council of Teachers of English, Champaign, Illinois, 1963, pp. 83-95.

11. Thompson, Wayne N., "A Study of the Grading Practices of Thirty-one Instructors in Freshman English," The Journal of Educational Research, 49: 65-68, 1955.

TWO-FACTOR EXPLANATION OF POST-HIGH

SCHOOL DESTINATIONS IN HAWAII[1]

PAUL W. DIXON
University of Hawaii, Hilo Campus

NOBUKO K. FUKUDA
University of Hawaii, Hilo Campus

ANNE E. BERENS
University of Hawaii

ABSTRACT

Data from 484 students out of a Hawaiian high school graduating class of 643 were used in a study of post-high school destinations. Of these 643 students, 360 were found to have selected academic destinations. They were going either to universities, a regional 2-year branch of the state university, 4-year colleges, junior colleges, or technical schools. A factor analysis of ten variables consisting of teachers' ratings of school related behavior, verbal (V) and quantitative (Q) School and College Abilities Tests (SCAT) scores, and rank in graduating class (R) revealed two factors: "g" a general intellective factor (SCAT Q and V) and school skills (teachers' ratings and R). In general the institutions of high academic quality attracted students who were significantly higher on the variables of both factors. Variation from this pattern of choice was discussed in terms of ego self-direction and social control as it affected mobility away from the island location.

ATTENTION should be given to the personality and academic characteristics of high school students, since they may affect students' choice of activity after high school. Students, counselors, and parents would find information on the characteristics variable valuable, since this would help them to predict the relationship between the students' performance in high school and his probable choice of destination after high school. Since post-high school destination choice is usually the first major decision influencing lifetime goals, it is important that the students and their advisors have this kind of information available.

The setting for this study is a "natural laboratory" variety, since the high school is located on one of the islands of the Hawaiian Island chain which has limited post-high school educational options. Students must make a relatively clear-cut decision between their home environment—with a particular kind of scholastic opportunity available—and a scholastic environment which imposes a geographical separation from their immediate home. The choice of post-high school destination is limited to attending a technical school or a 2-year campus of a state university if they remain on the home island or attending a university or 4-year college if they leave the island.

Seibel (9) found that about 40 to 49 percent of low ability (lower quarter of the distribution) high school seniors attending college choose 2-year institutions.

This is in direct contrast with the high ability (upper quarter) group where 90 percent attended 4-year institutions and only 10 percent attended junior colleges. Berdie (1) found that for students who plan to attend college mean American College Entrance (ACE) scores were higher than those of any other group. A greater proportion of this group chose attending college over all other alternatives. Measures of ability are thus predictive of choice of type of college (i. e., whether 2- or 4-year) and are also predictive of plans for college attendance.

A related study was performed by Richards and Braskamp (7) which analyzed the differences between students attending 2-year and 4-year colleges. Their findings indicated that students who attended 2-year campuses (1) tended to be less able academically than those attending 4-year colleges on both the aptitude (APT) test and on high school grade point averages; (2) varied more in academic talent than did students in 4-year colleges; and (3) had fewer nonacademic accomplishments except in art than did 4-year college students. Thus, "2-year colleges attract pragmatic students seeking vocational training; they are less attractive to talented students, or intellectually and academically oriented, who plan a degree in one of the traditional subject areas, and who expect to take part in a wide variety of activities in college." They concluded that junior-college students typically have different goals

FIGURE 1

TEACHERS' RATING SCALE

	0	10	20	30	40	50	60	70	80	90	100
ACCURACY (ACCU)		Work is poor. Makes frequent errors.		Work is inaccurate and below standard.			Work is well done and reasonably accurate.			Work is of highest quality.	
COOPERATION (COOP)		Disagreeable. Cannot or will not work with others.		Works with others sometimes but has difficulty.			Usually agreeable. Generally willing to help.			Always agreeable. Willing to do extra favors.	
EFFORT-INDUSTRY (E-I)		Does as little as possible. Lazy.		Seldom completes required work.			Usually does work that is required. Occasionally does extra work.			Very industrious. Does extra work gladly.	
INITIATIVE-LEADERSHIP (I-L)		Acts only under direction.		Seldom originates any work. Follows others.			Plans many of his activities and still needs supervision.			Marked ability to think for himself.	
RELIABILITY-RESPONSIBILITY (R-R)		Neglects promise and obligations. Unreliable.		Reliable on some occasions. Often needs supervision.			Usually dependable. Conscientious.			Thoroughly dependable.	
PROMPTNESS-PUNCTUALITY (P-P)		Undependable. Almost always late.		Frequently late.			Usually on time but occasionally late.			Always on time.	
SELF-CONFIDENCE (S-F)		Timid. Hesitant. Easily influenced.		Appears to be over self-conscious.			Wholesomely self-confident.			Shows superb self-assurance.	

than students attending 4-year colleges. The student attending a 2-year campus is more pragmatically oriented and desires more technical instruction in comparison to the student attending a 4-year campus who is looking for more academic, intellectual stimulation.

A factor analytic study by Richards and Holland (8) indicated that there were four major areas of influence which determine college choice: intellectual emphasis, practicality, advice of others, and social emphasis. They also found that these four areas are highly similar for men and women. The differences in college choice are related to aspects of the institution which the student is interested in attending.

A factor analysis of SCAT V and Q, teachers' ratings, and rank in high school class was performed in order to elucidate the simple structure among all variables. This study attempts to analyze the relationships which obtain between school-generated measures and post-high school destinations. Comparisons were made for all variables for every student to show the value of each. Comparisons were made between males (M) and females (F) who attended the 2-year branch of the university of Hawaii Hilo Campus on the island of Hawaii (UHHC), or who went to the main branch of the University on the island of Oahu (UHM). Comparisons were also made between those attending UHM and those attending colleges on the mainland United States (MNLD COLL). Comparisons were also made between students attending junior college on the mainland (MNLD JC) and those attending junior college in Hawaii (HAWAII JC). Similar comparisons were made between students attending college in Hawaii (HAWAII COLL) and colleges on the mainland (MNLD COLL). Overall comparisons were made between students attending college (COLL) and students attending junior college both in Hawaii and on the mainland (ALL JC). Another comparison was made between those attending the 2-year branch of UHHC and those attending junior college elsewhere (ALL JC). A similar comparison was made between those attending the 2-year campus of UHHC and those attending universities on the mainland. A similar comparison was made between those attending the 2-year UHHC campus and those attending colleges on the mainland. An overall comparison was made between technical school students (TECH) and all college and university students (ALL COLL). Another comparison was made between males attending technical school and males attending colleges and universities. A comparison was also made between females attending technical schools and females attending colleges and universities.

METHOD

Subjects

The students were members of the graduating class of a large Hawaiian high school. Those students who were selected for study had complete records on measures of SCAT, Teachers' Ratings, and Rank in Class (R). From a graduating class of 643, 484 were reached in the study the semester after June graduation. The scores of these students were used to obtain the factors by means of an oblique solution to factor rotation. Of those reached, 360 had post-high school destinations of an academic nature. The scores from these students were used in the compar-

isons between students attending universities, colleges, junior colleges, and technical schools. The instrumentation and procedure follow those described in an earlier publication (4).

RESULTS

The subroutines (PERSUB) of Bottenberg and Ward (2) and an IBM 7040 computer were used to determine the extent to which post-high school destination and sex variables contributed to differences in SCAT, Teachers' Ratings, and R. PERSUB used the statistical techniques of multiple linear regression to determine F ratios and exact probability values to 4-place accuracy. It presents an efficient method for programming computation of the estimated probability for any specific F value. The estimate is based on the actual distribution of scores and may be used with df ranging from 4 to 1,000 (6). The F ratio is computed between a full model regression equation containing all predictor variables under consideration and a restricted model.

In this analysis the information about a particular variable is not included in the restricted model's equation, and the predictive efficiency of this equation is compared with the predictive efficiency of the full model's equation, in which all the information for each variable is included (for complete discussion, see 4). For example, in predicting SCAT V scores the full model would include the female UHHC and female UHM differences in the population, while the restricted model omits the information that different destinations for females existed. If knowledge of destination differences helped in the prediction of SCAT V scores, there would be a significant difference using the F ratio statistic, between the full model equation's R^2 containing destination differences, and the restricted model's R^2 which does not include this information. The hypothesis tested is that members of the sample have different destinations but the same expected score on SCAT V.

Rotation was by means of Digman's (3) method, which is essentially a variant of the Harris-Kaiser system (5). In this type of factor analysis the diagonals are the squared multiple correlations on the remaining diagonals so that the values in the diagonals are the variance which these diagonals have in common with the other variables. Therefore, the values in the diagonals are based on the data rather than on the assumed value that the diagonals should have. A 10x10 variance-covariance matrix was rotated by this method.

The correlations among the ten factors are shown in Table 1; two factors were found to have an eigenvalue > 1.0. The correlation matrix was rotated by the Varimax procedure with the results shown in Table 2. This table presents the factor loadings according to an oblique solution. No allowance was made for differences between sexes in this analysis. This follows the work of Richards and Holland (8), who found no differences due to sex in their factor analysis of major influences in college choice from a 27-item questionnaire describing influences on college choice.

Table 3 shows the intercorrelations (factor cosines) between the three factors derived from the factor analysis. Table 3 shows that a high positive

TABLE 1

INTERCORRELATIONS BETWEEN SCAT V AND Q, TEACHERS' RATINGS, AND R

	1	2	3	4	5	6	7	8	9	10
1 = SCAT V	1.00	0.67	0.51	0.48	0.48	0.53	0.50	0.45	0.45	-0.56
2 = SCAT Q		1.00	0.54	0.51	0.52	0.52	0.55	0.47	0.42	-0.58
3 = ACCU			1.00	0.85	0.89	0.88	0.87	0.78	0.74	-0.77
4 = COOP				1.00	0.93	0.89	0.93	0.81	0.78	-0.72
5 = E - I					1.00	0.92	0.94	0.82	0.76	-0.76
6 = I - L						1.00	0.91	0.77	0.81	-0.74
7 = R - R							1.00	0.84	0.72	-0.76
8 = P - P								1.00	0.64	-0.67
9 = S - F									1.00	-0.59
10 = R										1.00

Note: Negative correlations between R (10) and other scores are the result of smaller R value showing higher rank in class.

relationship obtains between the two factors.

An analysis was performed using the statistical technique of multiple linear regression comparing students with different destinations after high school on the same ten variables. Table 4 shows the list of means for the comparisons between the 2-year branch of the university and students in other colleges and universities. It shows comparisons between students in the island university system and compares them with college students on the mainland (MNLD JC, and students attending HAWAII JC. A similar comparison is made between HAWAII COLL and MNLD COLL students. ALL COLL students are compared with ALL JC students. A comparison is also made between University of Hawaii students at the 2-year campus (UHHC) and those attending universities on the mainland (MNLD UNIV), and a similar comparison is made between students attending the 2-year campus of the university (UHHC) and those attending colleges on the mainland (MNLD COLL). An overall comparison is made between those attending technical school (TECH) and those attending all colleges and universities (COLL). This analysis compared males in colleges and universities (M ALL COLL) and males attending technical schools (MALE TECH). Similarly, female technical school students (FEMALE TECH) and females attending colleges and universities are compared (F ALL COLL). Table 5 presents all the comparisons listed here giving R-square, df, and F ratios.

DISCUSSION

Table 2 shows the two factors that were obtained from an oblique analysis of the correlation matrix

TABLE 2

OBLIQUE LOADINGS OF TEN VARIABLES ON TWO FACTORS

	School Skills 1	"g" 2
1 = SCAT V	-0.12	0.62 x
2 = SCAT Q	-0.06	0.57 x
3 = ACCU	0.48 x	0.07
4 = COOP	0.57 x	-0.03
5 = E - I	0.56 x	-0.02
6 = I - L	0.52 x	0.03
7 = R - R	0.53 x	0.01
8 = P - P	0.53 x	0.01
9 = S - F	0.53 x	-0.01
10 = R	-0.31 x	-0.26

x meets criterion ± .30

found in Table 1. Using the criterion that factor-loadings must be greater than ± .30 to be consider-ed meaningful, it was found that on Factor 1 teachers'

ratings and R load positively. (Since the scale for R is reversed, a negative value shows positive prediction.) Factor 2 appears to be a "g" or general intellective factor, since only SCAT V and Q meet the criterion of loading weight on that factor. It is easily described, since it shows intellective operations. Factor 1 may be described as a school skills variable, since teachers' ratings and R weigh positively on Factor 1.

TABLE 3

FACTOR CORRELATIONS

	School Skills 1	"g" 2
1	1.00	+0.84
2		1.00

Table 2 presents an interesting picture, since we see that teacher-related measures and R as defined by Factor 1 (school skills factor) are positively correlated with the "g" factor. This shows that teachers typically rate students high on all rating scales if they are high on "g." The reverse of this may also hold true.

Dixon and others (4) have shown that some teachers' ratings which are related to Factor 1 predict significantly to R as do the SCAT scores. Not all variables which meet the criterion on Factor 1 show significant independent contribution to prediction of R, however. Only accuracy (ACCU), Effort and Industry (E-I), and Reliability and Responsibility (R-R) make significant independent contribution to prediction of R. Thus, these three teacher rating variables are more salient in prediction of high school performance.

From the analysis of the differences in SCAT scores, teachers' ratings, and rank in class, we find a pattern of differences which show a rather uniform set of mean differences among individuals going to various post-high school destinations. Thus, for females attending the major university in the state as compared with females attending the 2-year-branch of the university, there are significant differences in the means for all variables showing that the former group has significantly higher scores on SCAT V and Q, teachers' ratings, and R. For males a somewhat different pattern can be seen, since males going to the main university campus are significantly higher on SCAT scores and rank in class but do not differ significantly from those attending the local branch of the university on teachers' ratings. Females in the first comparison would thus differ on both factors. All factors would contribute to differences in personality and intellectual profiles of females attending UHHC as compared with those attending the main university campus. Males differ only on Factor 2, the "g" factor, and not on the other factor. The ability level of a male student would be predictive of transfer to the main campus of the university, while high school performance would not be important. The significant difference in rank in class in favor of those attending the main university campus would therefore be due to differences in intellectual ability rather than in school-related personality characteristics.

Another comparison was made between students attending the main campus of the university and those attending college on the mainland. In this case the major determinant was a significant difference in rank in class in favor of those going to mainland universities. Intellectual ability and school-related personality characteristics would therefore not be important as far as post-high school destination is concerned in this instance. Actual performance in high school is, therefore, the best measure for determining post-high school destination for students who are leaving the immediate vicinity to go to colleges on the mainland.

A comparison was made between students attending junior colleges on the mainland and junior colleges in Hawaii, excluding the 2-year branch (UHHC). Here no significant differences were found. Another comparison was made between students attending HAWAII COLL and those attending MNLD COLL. Again no significant differences were obtained in this comparison. A further comparison was made between those attending COLL and those attending ALL JC excluding the 2-year branch (UHHC). In this comparison no significant differences were obtained. Thus, these students are not distinguishable to a significant degree on the measures which were obtained even though the means for the students attending 4-year colleges are in all cases above the means of the students attending junior colleges. The lack of significant difference here may be due to the small number of subjects available for these comparisons.

Students attending the 2-year regional branch (UHHC) were compared with students attending junior colleges in Hawaii or on the mainland. The result was that students attending the 2-year branch of UHHC had significantly higher mean scores on the variables included in both factors. The significant differences obtained might only indicate that admission standards differ between junior colleges and the 2-year campus of the state university. These differences might, on the other hand, reveal that a 2-year campus of a university attracts better students than do junior colleges which offer technical vocational education in addition to the college transfer program.

Similarly, a comparison between students attending the 2-year branch (UHHC) and those attending universities on the mainland revealed a uniform superiority on all variables in favor of students who attend mainland universities except for SCAT Q. Selection procedures at mainland universities seem to favor the more highly verbal student. The students at the 2-year UHHC did significantly better on SCAT Q. Quantitative abilities are thus related to lack of mobility toward mainland university attendance as compared with all other variables. This is the case even though SCAT V and Q are positively correlated to some degree ($r = +.67$). Quantitative abilities may therefore relate to a constellation of personality attributes which are negatively related to mobility. An

TABLE 4

MEANS OF ALL GROUPS

N		SCAT %		Teachers' Ratings							
		V	Q	ACCU	COOP	E - I	I - L	R - R	P - P	S - F	R
13	Female UHM	76.01**	83.92**	10.25**	10.54**	10.38**	9.99**	10.59**	11.14**	9.52**	102.85**
82	Female UHHC	72.40	82.49	9.86	10.44	10.02	9.84	10.44	10.97	9.18	137.16
12	Male UHM	66.19**	86.34**	9.42	10.37	9.99	9.48	10.29	10.93	9.22	175.25*
71	Male UHHC	66.42	85.00	8.97	9.48	9.21	8.84	9.67	10.08	8.82	223.92
25	UHM	71.29	85.08	9.85	10.44	10.20	9.74	10.45	11.04	9.38	137.60*
15	MNLD COLL	63.87	75.20	9.45	9.91	9.62	9.52	9.83	10.19	9.23	212.40
6	MNLD JC	26.33	44.33	8.15	9.10	8.60	8.08	8.78	9.90	7.95	352.67
7	HAWAII JC	33.86	77.17	7.93	8.81	8.38	7.08	8.64	10.03	7.51	326.86
6	HAWAII COLL	41.50	66.33	9.00	9.95	9.45	8.77	9.40	10.55	8.35	239.83
15	MNLD COLL	63.87	75.20	9.45	9.91	9.62	9.52	9.83	10.19	9.23	212.40
21	COLL	57.48	72.67	9.32	9.92	9.57	9.30	9.71	10.30	8.98	220.24
13	ALL JC	30.38	56.08	8.03	8.95	8.48	7.55	8.71	9.97	7.72	338.77
13	ALL JC	30.38**	56.08**	8.03**	8.95**	8.48**	7.55**	8.71**	9.97**	7.72**	338.77**
153	UHHC	69.62	83.65	9.44	10.16	9.65	9.38	10.08	10.55	9.01	242.78
153	UHHC	69.62**	83.65**	9.44**	10.16**	9.65**	9.38**	10.08**	10.55**	9.01**	242.78**
18	MNLD UNIV	77.70	83.47	9.79	10.62	10.02	10.12	10.45	10.73	9.85	212.40
153	UHHC	69.62**	83.65**	9.44**	10.16**	9.65**	9.38**	10.08**	10.55**	9.01**	242.78**
15	MNLD COLL	63.87	75.20	9.45	9.91	9.62	9.52	9.83	10.19	9.23	212.40
155	TECH	31.56**	53.50**	8.30**	9.18**	8.50**	8.06**	8.85**	9.77**	8.11**	358.44
189	COLL	57.68	72.67	9.32	9.92	9.57	9.30	9.71	10.30	8.98	220.24
77	Male TECH	33.42**	55.62**	8.01**	8.93**	8.18**	7.90**	8.42**	9.30**	7.90**	405.26**
111	M ALL COLL	52.64	68.64	8.88	9.37	8.99	8.67	9.12	9.66	8.48	289.18
119	Female TECH	29.72**	51.40**	8.58**	9.43**	8.83**	8.22**	9.28**	10.22**	8.32**	312.23**
78	F ALL COLL	62.80	77.10	9.81	10.52	10.21	10.00	10.36	10.99	9.52	144.40

* p < .05 level of significance.   **p < .01 level of significance.

TABLE 5

COMPARISON OF ALL GROUPS SHOWING R^2, df, F, and P VALUES FOR EACH COMPARISON

		R^2	df	F
R	FUHM UHHC	0.20	1/93	23.39**
SCAT V	FUHM UHHC	0.17	1/93	19.48**
SCAT Q	FUHM UHHC	0.10	1/93	10.36**
ACCU	FUHM UHHC	0.16	1/93	18.39**
COOP	FUHM UHHC	0.12	1/93	13.38**
E - I	FUHM UHHC	0.16	1/93	19.37**
I - L	FUHM UHHC	0.14	1/93	15.89**
R - R	FUHM UHHC	0.13	1/93	13.97**
P - P	FUHM UHHC	0.14	1/93	15.09**
SELF	FUHM UHHC	0.08	1/93	7.73**
R	MUHM UHHC	0.04	1/81	3.82*
SCAT V	MUHM UHHC	0.09	1/81	7.81**
SCAT Q	MUHM UHHC	0.11	1/81	10.49**
ACCU	MUHM UHHC	0.02	1/81	1.52
COOP	MUHM UHHC	0.03	1/81	2.26
E - I	MUHM UHHC	0.02	1/81	1.82
I - L	MUHM UHHC	0.02	1/81	1.82
R - R	MUHM UHHC	0.01	1/81	.97
P - P	MUHM UHHC	0.02	1/81	1.32
SELF	MUHM UHHC	0.02	1/81	1.95
R	UHM MNLD	0.09	1/41	3.95*
SCAT V	UHM MNLD	0.08	1/41	3.95
SCAT Q	UHM MNLD	0.05	1/41	2.17
ACCU	UHM MNLD	0.06	1/41	2.64
COOP	UHM MNLD	0.06	1/41	2.61
E - I	UHM MNLD	0.06	1/41	2.77
I - L	UHM MNLD	0.08	1/41	3.42
R - R	UHM MNLD	0.04	1/41	1.59
P - P	UHM MNLD	0.05	1/41	2.18
SELF	UHM MNLD	0.07	1/41	2.99
R	HAW MNL JC	0.00	1/11	0.01

TABLE 5 (Continued from previous page)

		R^2	df	F
SCAT V	HAW MNL JC	0.01	1/11	0.09
SCAT Q	HAW MNL JC	0.01	1/11	0.07
ACCU	HAW MNL JC	0.00	1/11	0.04
COOP	HAW MNL JC	0.00	1/11	0.04
E - I	HAW MNL JC	0.00	1/11	0.02
I - L	HAW MNL JC	0.01	1/11	0.11
R - R	HAW MNL JC	0.00	1/11	0.00
P - P	HAW MNL JC	0.00	1/11	0.00
SELF	HAW MNL JC	0.01	1/11	0.08
R	COL HAW MNL	0.01	1/19	0.20
SCAT V	COL HAW MNL	0.01	1/19	0.25
SCAT Q	COL HAW MNL	0.01	1/19	0.15
ACCU	COL HAW MNL	0.01	1/19	0.16
COOP	COL HAW MNL	0.01	1/19	0.10
E - I	COL HAW MNL	0.01	1/19	0.14
I - I	COL HAW MNL	0.01	1/19	0.18
R - R	COL HAW MNL	0.01	1/19	0.14
P - P	COL HAW MNL	0.01	1/19	0.12
SELF	COL HAW MNL	0.01	1/19	0.14
R	COLL ALL JC	0.01	1/32	0.35
SCAT V	COLL ALL JC	0.01	1/32	0.48
SCAT Q	COLL ALL JC	0.01	1/32	0.27
ACCU	COLL ALL JC	0.01	1/32	0.36
COOP	COLL ALL JC	0.01	1/32	0.26
E - I	COLL ALL JC	0.01	1/32	0.29
I - L	COLL ALL JC	0.02	1/32	0.51
R - R	COLL ALL JC	0.02	1/32	0.54
P - P	COLL ALL JC	0.01	1/32	0.19
SELF	COLL ALL JC	0.01	1/32	0.36
R	ALL JC UHHC	0.20	1/164	41.20**
SCAT V	ALL JC UHHC	0.25	1/164	55.91**
SCAT Q	ALL JC UHHC	0.21	1/164	44.16**

TABLE 5 (Continued from previous page)

		R^2	df	F
ACCU	ALL JC UHHC	0.13	1/164	24.42**
COOP	ALL JC UHHC	0.12	1/164	22.66**
E - I	ALL JC UHHC	0.12	1/164	23.50**
I - L	ALL JC UHHC	0.12	1/164	23.18**
R - R	ALL JC UHHC	0.22	1/164	38.86**
P - P	ALL JC UHHC	0.09	1/164	16.56
SELF	ALL JC UHHC	0.08	1/164	14.20**
R	UHHC MNLD UNIV	0.27	1/169	61.46**
SCAT V	UHHC MNLD UNIV	0.32	1/169	81.41**
SCAT Q	UHHC MNLD UNIV	0.25	1/169	56.40**
ACCU	UHHC MNLD UNIV	0.17	1/169	34.26**
COOP	UHHC MNLD UNIV	0.17	1/169	33.90**
E - I	UHHC MNLD UNIV	0.16	1/169	32.67**
I - L	UHHC MNLD UNIV	0.18	1/169	37.51**
R - R	UHHC MNLD UNIV	0.24	1/169	45.88**
P - P	UHHC MNLD UNIV	0.11	1/169	21.75
SELF	UHHC MNLD UNIV	0.13	1/169	25.95**
R	UHHC MNL COL	0.22	1/166	47.39**
SCAT V	UHHC MNL COL	0.28	1/166	65.42**
SCAT Q	UHHC MNL COL	0.23	1/166	50.33**
ACCU	UHHC MNL COL	0.14	1/166	28.30**
COOP	UHHC MNL COL	0.13	1/166	24.90**
E - I	UHHC MNL COL	0.14	1/166	26.76**
I - L	UHHC MNL COL	0.14	1/166	26.94**
R - R	UHHC MNL COL	0.07	1/166	12.76**
P - P	UHHC MNL COL	0.10	1/166	18.27**
SELF	UHHC MNL COL	0.09	1/166	16.54**
R	TECH ALL COLL	0.41	1/383	269.19**
SCAT V	TECH ALL COLL	0.40	1/383	255.36**
SCAT Q	TECH ALL COLL	0.34	1/383	194.35**
ACCU	TECH ALL COLL	0.28	1/383	152.66**

TABLE 5 (Continued from previous page)

		R^2	df	F
COOP	TECH ALL COLL	0.26	1/383	136.72**
E - I	TECH ALL COLL	0.28	1/383	147.93**
I - L	TECH ALL COLL	0.27	1/383	143.22**
R - R	TECH ALL COLL	0.32	1/383	160.83**
P - P	TECH ALL COLL	0.24	1/383	122.30**
SELF	TECH ALL COLL	0.18	1/383	81.67**
R	MTECH M ALL	0.23	1/186	54.34**
SCAT V	MTECH M ALL	0.21	1/186	48.91**
SCAT Q	MTECH M ALL	0.18	1/186	41.14**
ACCU	MTECH M ALL	0.14	1/186	31.21**
COOP	MTECH M ALL	0.13	1/186	27.68**
E - I	MTECH M ALL	0.14	1/186	30.70**
I - L	MTECH M ALL	0.13	1/186	28.46**
R - R	MTECH M ALL	0.41	1/186	87.11**
P - P	MTECH M ALL	0.13	1/186	28.21**
SELF	MTECH M ALL	0.10	1/186	21.09**
R	FTECH F ALL	0.24	1/195	60.94**
SCAT V	FTECH F ALL	0.24	1/195	62.96**
SCAT Q	FTECH F ALL	0.16	1/195	36.08**
ACCU	FTECH F ALL	0.18	1/195	44.12**
COOP	FTECH F ALL	0.16	1/195	36.56**
E - I	FTECH F ALL	0.19	1/195	47.05**
I - L	FTECH F ALL	0.18	1/195	43.10**
R - R	FTECH F ALL	0.26	1/195	60.78**
P - P	FTECH F ALL	0.18	1/195	42.25**
SELF	FTECH F ALL	0.10	1/195	22.69**

Note: Full model regression equation: General form based on binary sectors denoting sex group membership:
Male = Female predicting to criteria of interest.

** $p < .01$ level of significance.

* $p < .05$ level of significance.

F = Female.

M = Male.

earlier study (4) showed that teachers' ratings on E - I were related more closely to SCAT Q than to SCAT V.

A further comparison was made between the 2-year branch (U H H C)and mainland college attendance. In this instance, the students attending the 2-year campus (U H H C) scored significantly higher on SCAT scores and on teachers' ratings of cooperation (COOP), E - I, initiative and leadership (I - L), R - R, and Punctuality (P - P) as compared with students attending mainland universities. Students attending mainland colleges, on the other hand, scored significantly higher on ratings of ACCU, I - L, and self-confidence (S - F) were also significantly higher in rank in class. Students at the 2-year U H H C were therefore higher on Factor 2 ("g") than those attending mainland colleges, while those attending mainland colleges were higher on some variables of Factor 1 (school skills) except for COOP, E - I, R - R, and P - P, on which the 2-year U H H C students were significantly higher. The students attending the 2-year UHHC campus would thus be characterized as being significantly higher in intellectual abilities and on teachers' rating scale variables which measure socialization into a high school setting, while those students attending mainland colleges were characteristically higher on variables measuring ego self-direction, such as S - F, I - L, and ACCU (see Figure 1). The significantly higher mean rank in class in high school of students attending mainland colleges shows this same type of effectiveness within a high school setting. Thus, in this comparison, the more self-directing student with more ego strength tends to do better in high school and is more likely to attend a mainland college. The brighter and more socially conscious student, by comparison, stays near home at the 2-year UHHC campus. This particular comparison shows the relationship between the factor pattern and the choice of post-high school destination, but it also indicated that the teachers' rating scales are not all measuring the same personality characteristics.

The comparisons between technical school students and those attending colleges and universities revealed a uniform superiority on all variables in favor of those attending the more academic institution. Technical school students are therefore different from those attending college in school related skills as well as on general ability factors.

SUMMARY AND CONCLUSIONS

Two factors were obtained from the oblique solution to factor rotation using SCAT V and Q, teachers' ratings and R. These factors were (1) a combination of teachers' rating scales and R, which was labeled a "school skills" factor and (2) "g," or a general intellective factor with SCAT V and Q meeting the weighting criterion.

In general, higher scores on variables weighting on both factors differentiated between students in the following order: universities > colleges > a 2-year branch of the state university > junior colleges > technical schools. There were, however, some exceptions to this picture. Females attending the main university campus in the state differed on both factors, while males under the same comparisons only

differed on "g" and R. For those attending colleges on the mainland as compared with those attending the main university campus, a significantly higher mean on R was found for those going to the mainland. A comparison between the 2-year campus (U H H C)and junior colleges showed overall superiority for the 2-year campus of the university on all variables, due probably to differences in selection procedures. A comparison between mainland university students and those attending the 2-year UHHC campus showed the university students significantly higher on all variables except SCAT Q. This suggests that verbal, rather than quantitative abilities, have a greater effect on mainland university attendance.

The comparison between students attending mainland colleges and the 2-year UHHC campus revealed the most interesting relationship between the factor analysis and post-high school destination. Significantly higher scores or variables weighting on Factor 1, showing ego self-direction in higher mean scores on ACCU, S - F, and I - L and higher R characterized students attending mainland colleges. Students attending the 2-year campus (U H H C)were significantly higher on Factor 2, "g," and teachers' rating of COOP, R - R, P - P, and E - I from Factor 1, and lower on R. Students attending mainland colleges are judged by teachers as being more self-confident and ego self-directed; furthermore, they are found to be more effective in high school, having a significantly higher mean R, while those attending the 2-year UHHC campus have high "g" but are more submissive, conscientious, and socially controlled. They are, thus, more likely to remain under parental control. High "g" does not necessarily make a student more ego self-directed and more venturesome; instead, it may relate to a greater degree of socialization.

The high positive correlation shown in Table 3 between Factor 1 (school skills) and Factor 2 ("g," overall intellectual ability) indicate that within this particular high school setting, ability, as measured by the SCAT, is predictive of greater competence in meeting the requirements of the school system.

FOOTNOTE

REFERENCES

1. Berdie, R. F. , "Why Don't They go to College?" Personnel and Guidance Journal, 31;(no. 6)352-356, 1953.

2. Bottenberg, P. A. ; Ward, J. H. , Applied Multiple

Linear Regression, Technical Documentary Report PRL-TDR-63-6, Defense Documentation Center, Defense Supply Agency, Washington, D.C., March 1963.

3.   Digman, J.M., "The Procrustes Class of Factor Analytic Transformations," Multivariate Behavioral Research, 2:89-94, 1967.

4.   Dixon, P.W.; Fukuda, Nobuko, E.; Berens, Anne E., "Effectiveness of Teachers' Ratings, Sex, and SCAT Scores as Predictors of Rank in High School Class," The Journal of Experimental Education, 37:(no. 3), 21-26, 1969.

5.   Harris, D.W.; Kaiser, H.F., "Oblique Factor Analytic Solutions," Psychometrika, 29: 347-362, 1964.

6.   Neel, G.I., Estimation of Probabilities Associated With the F Statistic by Digital Computer Techniques, Technical Documentary Report PRL-TDR-63-7, Defense Documentation Center, Defense Supply Agency, Washington, D.C., March 1963.

7.   Richards, J.M., Jr.; Braskamp, L.A., "ACT Studies 'Who Goes Where to Junior College,'" Activity, 6:(no. 1), 1967.

8.   Richards, J.M., Jr.; Holland, J.L., "A Factor Analysis of Student 'Explanation' of Their Choice of a College," ACT Research Reports, No. 8, 1968.

9.   Seibel, D.W., "The Relationships of Some Academic Ability Characteristics of High School Senior to College Attendance and Performance," College and University, 42(no. 1)41-52, 1966.

BOOK   REVIEWS

Robert   E.   Clasen   book   review   editor

BEHAVIORAL PROBLEM CHILDREN IN THE SCHOOLS

Woody, Robert H., (New York: Appleton-Century-Crofts, 1969), 264 pp.

BEHAVIORAL Problem Children in the Schools attempts to give classroom teachers, school counselors, and school psychologists an understanding of behavior-problem children and an approach to meet the needs of these children. This approach, which combines learning theory and conditioning techniques with insight-oriented techniques of counseling and psychotherapy, the author calls psycho-behavioral therapy. Although these theoretical positions differ, Woody feels their common elements used conjointly constitute the most effective means of meeting the needs of most behavioral-problem children.

The behavioral-problem child is defined as "the child who cannot or will not adjust to the socially acceptable norms for behavior and consequently disrupts his own academic progress, the learning efforts of his classmates, and interpersonal relations." Woody feels that "every child has, or could conceivably have, behavior problems at some point in his life and could, therefore, be considered a behavioral-problem child." The reviewer would disagree with Woody about applying a label to so large a group of children, particularly in view of Woody's own statements in chapter two in which he points out that what is a behavior problem in one situation may not be in a different situation. Few that the personal opinions, theoretical and professional orientation of the observer will influence the definition. Few would argue, however, with his premise that it is urgent for educators to learn how to meet the needs of these children.

In part one, Woody outlines the causes and characteristics of behavior problems and covers detection, referral, and psycho-educational diagnosis. To carry out an adequate psycho-educational diagnosis, Woody feels that tests must be administered, scored, and results entered; the diagnostician must give a clinical opinion as to the causes of the problem; and it must be determined what can be expected of the child and what can be done to help him. "Psycho-educational diagnosis should involve more than one diagnostic technique" and those that Woody discusses as making important contributions are the social case history, the psychological survey, the psychiatric, the neurological, and the pediatric examination.

In part two, the author discusses various approaches to behavioral change such as guidance, counseling, psychotherapy, and behavior therapy. He does not claim that the position he advocates, namely psycho-behavioral therapy, is a new theory, but rather an outline of practical techniques that have been gleaned from the various approaches that can be used to change behaviors. He states that not enough is known about individual differences and the validity of all aspects of the different approaches to claim that one specific approach has universal applicability. Advocates of the diverse theoretical positions would find it difficult to disagree with this, but would undoubtedly be opposed to Woody's statement that "What is perhaps more significant than the practical aspects of combining insight and action systems is the fact that behavior therapy and the counseling-psychotherapy

THE EFFECT OF IMAGE SIZE

ON VISUAL LEARNING

FRANCIS M. DWYER
The Pennsylvania State University

ABSTRACT

The purpose of this study was to investigate the effectiveness of four types of visual illustrations used to complement oral instruction and to compare their relative effectiveness when projected on viewing areas of different sizes. Each of the 588 Ss received a pretest, participated in his respective presentation, and received four individual criterial measures. Results indicated that (a) the use of illustrations to complement oral instruction does not automatically improve achievement, and (b) merely increasing the size of visual images used to complement oral instruction will not necessarily improve achievement.

IN A PREVIOUS paper (1) the author reported the results of a study designed to measure the effectiveness with which varied types of visual illustrations facilitate student achievement of different educational objectives via the medium of television using 22-inch monitors. The results demonstrated that the use of visual illustrations to complement oral instruction presented to college students does not necessarily improve their achievement on tests measuring different objectives. Only on one criterial measure was the use of visuals found to be an important instructional variable in increasing student achievement.

The purpose of the present evaluation was to replicate and extend the cited study. The purposes of the present evaluation were to determine whether:(a) the same results that occurred when students received their visualized presentation via 22-inch monitors would occur when they received the same instruction by means of 5-by-3 foot front projection images and by 6-by-4 foot rear projection, and (b) the same visuals presented in different sizes (i. e., 1½-by-1 foot, 5-by-3 foot, 6-by-4 foot) would be equally effective in facilitating student achievement of different educational objectives.

PROCEDURE

The content material for this evaluation was a 2,000-word instructional unit discussing the human heart, its parts, and internal operations. The taped instruction also contained audio signals which cued the change of slides so that the appropriate visuals appeared simultaneously with the oral instruction they were designed to complement.

Each experimental treatment was complemented by thirty-nine black and white slides. These slides were specially designed to illustrate a specific item of information presented in the oral instruction. Slides used in each sequence displayed relatively the same information, differing only in the amount of detail they contained. Each instructional presentation was video taped on an Ampex 660B video tape recorder.

Students in each of the three studies received the same instructional treatments; however, the method of presentation differed for each. Students in Study I, the control study, received their instruction in conventional television classrooms via 22-inch monitors; students in Study II received instruction by means of a Telebeam Model A-912-A television projector which provided a 5-by-3 foot projected image, and students in Study III received instruction by means of a Telebeam Model A-912-A television projector which projected a 6-by-4 foot rear screen image.

TREATMENT GROUPS

Speech 200 classes at The Pennsylvania University supplied the 588 Ss for this evaluation. Students in each of the studies were assigned to treatment groups according to which of the five instructional sessions they would be able to attend. The

experimental treatments were assigned to the five groups at random (see Table 1).

TABLE 1

NUMBERS OF STUDENTS IN EACH TREATMENT FOR EACH STUDY

Treatment	Study I	Study II	Study III
Oral Presentation (Group I)	62	35	32
Drawing Presentation (Group II)	54	35	29
Detailed Shaded Presentation (Group III)	54	30	36
Heart Model Presentation (Group IV)	51	31	30
Photographic Presentation (Group V)	48	31	30

In each study students in Group I, the control group, received no illustrations of the heart, but they viewed slides containing the names of the parts and processes of the heart as they were mentioned orally. Group II viewed simple line illustrations of the heart. Group III viewed detailed, shaded drawings of the heart. Group IV viewed photographs of a heart model, and Group V viewed realistic heart photographs. All groups received the same oral instruction and viewed their respective instructional presentation for equal amounts of time.

CRITERIAL MEASURES

In each study each student in each treatment group received the Otis Quick Scoring Mental Ability Test as a pretest, participated in his respective instructional presentation, and then received four individual criterial tests. Scores received on these tests were combined into a 78-item total criterial test. The objective of each test was as follows: (a) the drawing test evaluated learning of specific locations of the patterns and positions of the parts of the heart; (b) the identification test measured transfer of learning, i.e., the ability to identify numbered parts on a diagram of the heart from information received in the instruction; (c) the terminology test evaluated knowledge of referents for specific symbols; (d) the comprehension test measured understanding of the heart, its parts, and its internal operations, and (e) the total criterial test measured the students total understanding of the concepts presented (see Table 2).

RESULTS

In each study the Hartley Test for homogeneity of variance (3:94-95) was used on scores achieved on the Otis Quick Scoring Mental Ability Test for the five treatment groups. In no case did the observed value of the F max statistic reach the critical value

TABLE 2

KUDER-RICHARDSON FORMULA 20 RELIABILITY COEFFICIENTS FOR THE FIVE CRITERIAL MEASURES

Criterial Tests	Study I (22-inch monitor)	Study II (5-by-3 foot front projection)	Study III 6-by-4 foot rear projection)
Drawing Test	.82	.84	.81
Identification Test	.79	.81	.76
Terminology Test	.85	.81	.83
Comprehension Test	.74	.76	.79
Total Criterial Test	.91	.93	.92

for a .05 level test. Thus, it appeared that the treatment groups in each study were drawn randomly from populations with common variance. The F-ratios resulting from the analysis of variance on scores achieved on the criterial tests indicated that significant differences existed among the means of the five treatment groups on the drawing test in each of the three studies (Study I: $F = 7.38$, $df = 4/264$, $p < .01$; Study II: $F = 3.32$, $df = 4/157$, $p < .05$; Study III: $F = 3.86$, $df = 4/152$, $p < .01$). In the three studies no significant differences existed among the means of the remaining four criterial tests.

For each study comparisons among the individual means of the five treatment groups on the drawing test were conducted via Tukey's W Procedure (2:344-345). In each of the three studies the oral presentation without visuals was as effective as the visually complemented treatments on four of the five criterial tests. The exception was the drawing test, for which in all three studies the abstract line presentation (Group II) was significantly more effective than the oral presentation without visuals (Group I) in facilitating student achievement (Study I: Group II > Group I, $W = 3.04$, $n/v = 5/269$, $p < .01$; Study II: Group II > Group I, $W = 2.80$, $n/v = 5/162$, $p < .05$; Study III: Group II > Group I, $W = 3.13$, $n/v = 5/157$, $p < .05$).

As was previously stated the second objective of this evaluation was to measure the effectiveness of identical visuals of oral instruction complemented by identical visuals of different sizes. Study I used $1\frac{1}{2}$-by-1 foot images presented by conventional 22-inch monitors; Study II used 5-by-3 foot front projection images; and Study III used 6-by-4 foot rear projection images. Analysis of variance was conducted on scores achieved on the Otis Quick Scoring Mental Ability Test across the three studies, comparing the groups receiving the same instructional treatment in each of the three studies. Results indicated that students in the equivalent group in each of the studies could be considered to have been drawn randomly from populations with common

TABLE 3

ANALYSIS OF THE EFFECTIVENESS OF THREE METHODS OF INSTRUCTIONAL PRESENTATION

Criterial Tests	Instructional Treatments				
	Oral Presentation (Group I)	Abstract Line (Group II)	Drawing Presentation (Group III)	Heart Model Presentation (Group IV)	Photographic Presentation (Group V)
Pretest	n. s.	n. s.	n. s.	n. s.	n. s.
Drawing	.01	.05	n. s.	n. s.	n. s.
Identification	.05	.05	.05	.05	n. s.
Terminology	n. s.	n. s.	n. s.	n. s.	n. s.
Comprehension	n. s.	n. s.	n. s.	n. s.	n. s.
Total Criterial	n. s.	.05	n. s.	n. s.	n. s.

variance. Analysis of variance was also conducted on scores achieved on each criterial test by students receiving the same instructional treatment in each study where significant differences were found to exist (see Table 3). Tukey's W-Procedure was used to measure the differences between pairs of means.

An analysis of variance indicated that for students who received the oral presentation significant differences were found to exist on the drawing (F = 5.83, df = 2/126, p < .01) and identification tests (F = 3.74, df = 2/126, p < .05). In analyzing the differences between pairs of means on the drawing test (Study I > Study III, W = 2.67, n/v = 3/129, p < .01) and identification test (Study I > Study III, W = 2.22, n/v = 3/129, p < .05), instruction complemented by $1\frac{1}{2}$-by-1 foot images was more effective than instruction complemented by 6-by-4 foot images (see Table 4).

For students receiving the abstract line presentation the analysis indicated that significant differences in mean achievement existed among the three methods of presentation on the drawing (F = 3.47, df = 2/115, p < .05), identification (F = 4.01, df = 2/115, p < .05), and total criterial test (F = 3.49, df = 2/115, p < .05). An analysis of the differences between pairs of means on the drawing and identification tests indicated that instruction complemented by $1\frac{1}{2}$-by-1 foot images was more effective than instruction complemented via 6-by-4 foot rear screen images (Study I > Study III, W = 2.38, n/v = 3/118, p < .05) and 5-by-3 foot front projection images respectively (Study I > Study II, W = 2.30, n/v/ = 3/118, p < .05). Differences between the method means on the total criterial test approached but did not reach the critical value necessary for significance at the .05 level (see Table 5).

For the detailed, shaded drawing presentation, significant differences were found to exist among the three methods of presentation on the identification test (F = 3.62, df = 2/117, p < .05). Analysis again indicated that instruction complemented by $1\frac{1}{2}$-by-1 foot images presented via 22-inch television monitors was more effective than instruction complemented by

6-by-4 foot rear screen images (Study I > Study III, W = 2.20, n/v = 3/120, p < .05) (see Table 6).

For the heart model presentation significant differences were found to exist among the three methods of presentation on the identification test (F = 3.16, df = 2/109, p < .05). Instruction presented to students via the 22-inch monitors was found to be more effective than oral instruction complemented by 5-by-3 foot front projection images (see Table 7).

DISCUSSION

In each of the three cited studies the oral presentation without visuals was as effective as the visually complemented treatments on four of the five criterial tests. The exception was the drawing test for which in all three studies the abstract line presentation was significantly more effective than the oral presentation without visuals in facilitating student achievement. These results tend to add confidence to the explanations presented in a previous study (1:40-41):

(a) Since college students are generally selected from the upper two-thirds of the population in terms of verbal and conceptual ability, it seems that they are in a highly favorable position in terms of being able to learn from oral instruction. If this assumption is accurate, then the use of visual illustrations is not necessary to complement oral instruction designed to promote learning objectives similar to those measured by the identification, terminology, comprehension, and total criterial test.

(b) The realistic detail contained within the visual illustrations used to complement the oral instruction may have had the net effect of distracting the attention of the students from the essential learning cues, thereby interfering with rather than facilitating student achievement.

(c) Since students in each treatment group

TABLE 4

TUKEY'S W-PROCEDURE FOR DIFFERENCES BETWEEN STUDY MEANS OF GROUP I: THE ORAL PRESENTATION

A. Drawing Test

					W- Values	
	N	SD	Mean IQ Score	Mean	Study II 9.37	Study III 8.22
Study I: 22-inch Monitors	62	2.9	122.07	10.89	1.52	2.67**
Study II: 5-by-3 foot Front Projection	35	3.8	122.14	9.37		1.15
Study III: 6-by-4 foot Rear Projection	32	4.9	121.59	8.22 '		

B. Identification Test

			W-Values	
	SD	Mean	Study II 12.86	Study III 11.17
Study I: 22-inch Monitors	3.4	13.39	.53	2.22*
Study II: 5-by-3 foot Front Projection	3.7	12.86		1.69
Study III: 6-by-4 foot Rear Projection	4.5	11.17		

* p < .05
** p < .01

TABLE 5

TUKEY'S W-PROCEDURE FOR DIFFERENCES BETWEEN STUDY MEANS OF GROUP II: ABSTRACT LINE PRESENTATION

A. Drawing Test

					W-Values	
	N	SD	Mean IQ Score	Mean	Study II 12.17	Study III 11.35
Study I: 22-inch Monitors	54	3.9	122.15	13.72	1.55	2.37*
Study II: 5-by-3 foot Front Projection	35	4.0	121.29	12.17		.82
Study III: 6-by-4 foot Rear Projection	29	4.7	122.35	11.35		

B. Identification Test

			W-Values	
	SD	Mean	Study II 12.26	Study III 13.10
Study I: 22-inch Monitors	3.7	14.56	2.30*	1.46
Study II: 5-by-3 foot Front Projection	4.0	12.26		.84
Study III: 6-by-4 foot Rear Projection	3.7	13.10		

* p < .05

TABLE 6

TUKEY'S W-PROCEDURE FOR DIFFERENCES BETWEEN STUDY MEANS OF GROUP III:
DETAILED, SHADED DRAWING PRESENTATION

A. Identification Test

					W-Values	
					Study II 13. 47	Study III 11. 86
	N	SD	Mean IQ Score	Mean		
Study I: 22-inch Monitors	54	4. 0	120. 19	14. 06	. 59	2. 20*
Study II: 5-by-3 foot Front Projection	30	3. 7	120. 43	13. 47		1. 61
Study III: 6-by-4 foot Rear Projection	36	3. 7	122. 83	11. 86		

* p < .05

TABLE 7

TUKEY'S W-PROCEDURE FOR DIFFERENCES BETWEEN STUDY MEANS OF GROUP IV:
HEART MODEL PRESENTATION

A. Identification Test

					W-Values	
					Study II 12. 36	Study III 12. 86
	N	SD	Mean IQ Score	Mean		
Study I: 22-inch Monitors	51	3. 3	120. 43	14. 29	1. 93*	1. 43
Study II: 5-by-3 foot Front Projection	31	4. 1	120. 03	12. 36		. 50
Study III: 6-by-4 foot Rear Projection	30	3. 6	119. 57	12. 86		

* p < .05

viewed their respective televised presentation for equal amounts of time, those students who viewed the more realistic types of visuals may not have had sufficient time to study and comprehend adequately the additional information contained in the visual illustrations presented to them.

The results obtained in evaluating the effectiveness of oral instruction complemented by visual images of different sizes were quite interesting. The data indicated that merely increasing the size of instructional illustrations by projecting them on larger viewing areas does not automatically improve their effectiveness. In fact, for certain learning objectives the use of the larger images inhibited student achievement. The results indicated that where significant differences occurred the instruction presented via the 22-inch monitors was most effective in promoting student achievement.

The success of the instruction presented on the 22-inch monitors may be explained by the fact that the visuals presented more clearly the information needed by students to achieve specific objectives.

When these same visuals were expanded on the larger screens, a deterioration of the picture quality resulted. The distinctive characteristics of the visuals appeared blurry making it increasingly difficult for students to perceive the intended messages.

Another possible explanation may be suggested to account for the effectiveness of instruction presented by means of the conventional television monitors—the increased size of the visual images produced a larger viewing area which required the students to spend more time searching for the relevant visual information being discussed orally. Apparently the ability to be able to perceive clearly the relevant instructional characteristics in visuals is prerequisite for visual learning.

SUMMARY

A number of important generalizations can be developed from the cited studies which may be helpful in guiding the production and use of visual illustrations used for instructional purposes on television.

1. The use of visual illustrations to

complement televised instruction does not automatically improve student achievement of different types of educational objectives.

2. The type of visual illustration found to be most effective in facilitating student achievement of a specific educational objective depends on the type of information needed by the student to achieve that objective.

3. Merely increasing the size of visual images used to complement television instruction will not necessarily improve student achievement.

REFERENCES

1. Dwyer, F.M., "When Visuals are not the Message," Educational Broadcasting Review, 2: 38-43, 1968.

2. Sparks, J.M., "Expository Notes on the Problem of Making Multiple Comparisons in a Completely Randomized Design," The Journal of Experimental Education, 31:343-349, 1963.

3. Winer, B.J., Statistical Principles in Experimental Design, McGraw-Hill Book Company, Inc., New York, 1962.

TEAM TEACHING, STUDENT ACHIEVEMENT,

AND ATTITUDES

NEAL R. GAMSKY
Waupun, Wisconsin

ABSTRACT

This study examined the effects of team teaching on selected attitudes and achievement of ninth grade students in English and World History after 1 year. Seventy-four ninth grade students from a local school were randomly assigned to a team teaching treatment group while the remaining seventy-one members of the freshman class received traditional instruction in three separate classrooms. A teacher constructed test was employed to measure achievement in subject matter, and a personality inventory was used to measure changes in selected attitudes. The findings indicated that the team teaching approach did not appear to complement academic growth over traditional teaching methods but it did have a significant impact on student attitudes toward teachers, interest in subject matter, sense of personal freedom, and self-reliance.

THIS IS a report of a pilot study in a larger investigation of the impact of team teaching on the attitudes and achievement of high school students.

Several investigations have compared team teaching to student achievement with varying degrees of success (2, 3, 5, 6, 9). The educational results of this type of cooperative teaching upon student achievement has not yet been clearly established. Most researchers have continued to rely upon standardized tests to gauge achievement even though it is doubtful that most experimental treatments introduced are of a nature to produce differences in outcomes on these measures (7). It is believed that the instrumentation should be specially suited to the treatment under study. Also, it is questionable whether the effects of team teaching on student achievement constitute the most appropriate area for inquiry (8). More research relating team teaching to variables other than achievement is needed (1). Heathers (4), in a review of research, noted the lack of studies relating specific program aspects (teacher demonstration, team leadership, flexible grouping, etc.) to outcomes.

PURPOSE

This study sought to explore the effect of team teaching on selected attitudes and achievement of ninth grade students in English and World History after 1 year using teacher constructed instruments specifically related to the subject matter taught. The same teachers employed the same course of study to team-taught, flexibly grouped students and to conventionally taught students. The specific hypotheses investigated were:

1. Ninth grade students in the team teaching program will perform significantly better on teacher constructed tests in English and History than students who are taught by traditional methods.

2. Ninth grade students who are team-taught will have significantly better attitudes toward self, school, course work, teachers, and class, than students in regular classrooms.

METHOD AND PROCEDURES

Seventy-four ninth grade students from a local school were randomly assigned to a team teaching treatment group while the remaining seventy-one members of the freshman class received traditional instruction in three self-contained classrooms. The team teaching program consisted of World History and English during a scheduled 2-hour period from 9:00-11:00 A. M. each day. The team was composed of an English teacher, a History teacher and a para-professional. Assistance was also received from the librarian for library research and a secretary for

TABLE 1

WITHIN AND BETWEEN TREATMENT GAIN SCORE COMPARISONS OF STUDENTS ON WORLD HISTORY AND ENGLISH ACHIEVEMENT TESTS IN TEAM-TAUGHT AND CONVENTIONAL CLASSES

		N	D	ϵD^2	$\frac{D}{N}$	$\overline{D}_{X-C}$	Within Group t Pre- Post	Between Group t X-C
WORLD HISTORY								
Semester I	X	74	2,331	78,237	31.50	3.81	33.38*	2.50*
	C	71	1,966	61,343	27.69		23.49*	
Semester II	X	73	1,730	49,130	23.69	3.01	.2098	1.49
	C	71	1,896	62,980	26.70		.1513	
ENGLISH								
Semester I	X	74	1,655	42,612	22.36	2.04	21.96*	1.29
	C	68	1,382	34,674	20.32		16.90*	
Semester II	X	73	2,538	96,550	34.76	.34	27.64*	.117
	C	68	2,341	90,183	34.42		23.72*	

*P< .05

X = Experimental Group
C = Conventional Group

clerical work. The program utilized the concept of flexible modular scheduling to provide for large group instruction, small group discussions, and independent study. Two (20 minute) modules were designated for large group instruction and four modules for small group activities including two modules for seminars, one module for library, and one module for independent study. These units of time were arranged in varying patterns to allow for individual differences in ability and interest. Although both teachers shared in the planning and preparation of lesson units, one teacher usually had responsibility for the presentation during the large group instructional period. Both teachers participated in the small group sessions. Supplementary assistance was provided each day during the 2-hour period by the paraprofessional who helped in the preparation of materials, grading of papers, and independent study. The course of study and the teaching staff were the same for the experimental and control groups.

Each teacher constructed a 100-item test to measure the subject matter covered in each semester in ninth grade English and Social Studies. The test, which consisted of true-false, multiple-choice, and completion items, was designed to be completed in one hour. Frequency of correct answers was used as a score. Both experimental and control groups received pretests and posttests in English and History for first and second semesters.

A correlated t-test was used to measure change resulting from the instructional approach in each group. The mean difference between pretests and posttests was measured to compare gain made by the experimental group with that made by the control group. Pretests in English and World History received the F-test for homogeneity of variance. Variances were not significantly different.

Attitude scale items were constructed by a team of psychologists and reviewed by the teachers who agreed that the attitudes listed were important and could be affected by team teaching. The attitude scale items, drawn from several sources such as common tests of personality and school attitude inventories by a 3-member panel of educators, were tried out in a group of thirty summer school students. The final hundred items were scored on a one to five agree-disagree scale, and actual responses were subtracted from the ideal response thereby giving a quantitative measure of distance from the ideal attitude. The difference between actual and ideal scores for pretests and posttests were expressed as absolute values. Gain scores were measured by comparing pretest and posttest differences between actual and ideal scores with a correlated t-test. The assumption of homogeneity of variance was tested on pretest total scores as well as on subtest scores. On no score did the F-test yield an observed value (equal to or greater than the critical value), the variances were not significantly different.

Prior to analysis, pretest responses were factor analyzed by centroid extraction and varimax rotation using an orthogonal solution. Mean variances were compared. Responses ranging from strongly agree to strongly disagree were used to analyze the hundred items in ten factor clusters based on communalities. Criteria for cessation of factor extraction was based on the amount of variance accounted for by the factors as well as by the relative loadings of variables in each factor. The hundred items identified with the following subscales:

1. Self-concept of ability
2. Self-discipline in study habits
3. Attitude toward school
4. Self-reliance (ability to act independently)

TABLE 2

WITHIN AND BETWEEN GROUP ATTITUDE GAIN SCORE COMPARISONS OF STUDENTS IN TEAM-TAUGHT AND CONVENTIONAL CLASSES

Subtest		Pretest		Posttest		Within Group X̄ Pre- Post Difference	t	Between Group X̄ Pre- Post Difference	t
		$\overline{X}$	sd	$\overline{X}$	sd				
1. Self-concept of ability	X	16.30	7.747	15.14	7.661	1.155	1.571	6.1960	
	C	18.39	7.217	17.44	8.539	.953	1.265	6.0275	.19140
2. Study habits - self-discipline	X	13.48	5.438	13.38	6.519	.098	.154	5.3696	
	C	15.53	6.294	16.89	7.299	-1.359	1.984*	5.4812	1.5598
3. Attitude toward school	X	10.28	4.574	10.49	4.988	-.211	.318	5.5957	
	C	11.66	5.289	11.94	5.594	-.281	.414	5.4291	7.3600
4. Self-reliance (ability to act independently)	X	12.45	5.887	10.54	5.472	1.915	3.337*	4.8366	
	C	13.98	5.397	12.64	5.533	1.344	2.500**	4.2992	.7227
5. Socialization	X	9.66	4.840	9.07	5.851	.595	.785	6.3462	
	C	9.391	5.383	10.19	6.192	-.796	1.173	5.4370	1.3577
6. Attitudes toward teachers	X	11.90	5.052	12.13	6.500	-.225	.264	7.1797	
	C	12.72	6.222	15.53	7.848	-2.813	3.396**	6.6258	2.1681*
7. Interest in subject matter	X	15.44	5.458	15.20	6.371	.239	.300	6.7241	
	C	17.59	6.438	19.45	6.563	-1.859	2.748**	5.4127	1.9838*
8. Class participation	X	14.41	6.815	14.42	7.007	-.014	.019	6.2140	
	C	16.48	7.028	15.44	8.657	1.047	1.131	7.4075	.9045
9. Respect for the opinion of others	X	9.183	4.966	8.54	5.571	.633	.948	5.6347	
	C	9.969	6.409	9.54	5.526	.421	.767	4.3999	.2417
10. Sense of personal freedom	X	14.27	6.031	12.59	6.665	1.676	2.164*	6.5264	
	C	15.73	6.717	16.55	8.002	-.812	1.095	5.9385	2.3083*
Composite	X	127.4	37.26	121.5	43.45	5.859	1.343	36.7638	
	C	141.5	44.17	145.6	48.54	-4.156	1.148	28.9588	1.7452***

*P< .05   **P< .01   ***P< .07

N_C = 64 (Conventional Group)

N_X = 71 (Experimental Group)

5. Socialization
6. Attitudes toward teachers
7. Interest in subject matter
8. Class participation
9. Respect for the opinion of others
10. Sense of personal freedom

RESULTS

The results of this study are presented in Tables

1 and 2. The findings from an analysis of the data in Table 1 can be summarized as follows:

1. Both groups made significant gains by the end of semester I in English and World History, but the experimental group gains in World History were significantly greater than those of the traditionally taught group.

2. There were no significant differences between groups for semester I English gains.

3. Semester II World History scores showed no significant differences within or between groups.

4. Semester II English scores showed significant gain within both groups, but none between groups.

Analysis of the data in Table 2 revealed the following major findings:

1. Measurement of within group changes showed that the traditionally taught group experienced significant improvement in attitudes on subtest 4 (Self-reliance) but significantly negative growth in attitudes on subtests 2 (Study habits), 6 (Attitudes toward teachers), and 7 (Interest in subject matter).

2. Within group analysis of the experimental group revealed significantly positive growth in attitudes on subtests 4 (Self-reliance) and 10 (Sense of personal freedom).

3. In comparing pretest and posttest mean differences between groups, the experimental group experienced significantly greater positive attitude growth in subtests 6 (Attitudes toward teachers), 7 (Interest in subject matter), and 10 (Sense of personal freedom).

CONCLUSIONS AND IMPLICATIONS

The findings tend to support those of Zweibelson (9). Team teaching apparently has little impact over more traditional approaches on student achievement in high school English or Social Studies when teacher made tests are used to reflect the specific subject matter taught. Significant differences, however, were obtained with regard to student attitudes. The team-taught students displayed greater growth in their feelings of self-reliance and personal freedom, had more positive attitudes toward the teacher and were more interested in the subject matter than the traditionally taught students.

In spite of the fact that there were positive experimental group gains (.07) on the total attitude scale score, some attitudes were unaffected. For example, one common agreement for team teaching has been that small discussion groups allow for greater class participation and interaction among students. Yet the data indicated that the traditionally taught groups felt they had the opportunity for greater participation in class than team-taught students. In view of the possible long range significance for behavior, further investigation would appear to be warranted with regard to the impact of organizational and interactional patterns on student attitudes.

Generalization from this study is limited due to the small number of students and teachers involved and the fixed nature of the school setting. However, it would seem that since there are specific benefits in student attitudes to be gained from participation in team teaching programs, school officials would be well advised to "sell" these programs to the public on that basis rather than on the unsubstantiated

grounds that students learn more subject matter from this method of teaching.

REFERENCES

1. Anderson, Robert, <u>Teaching in a World of Change</u>, Harcourt, Brace, and World, New York, 1966.

2. Bair, Medill; Woodward, Richard, <u>Team Teaching in Action</u>, Houghton Mifflin Co., Boston, Massachusetts, 1964.

3. Georgiades, William; Bjelke, Joan, "An Experiment in 5-Day-A-Week Versus 3-Day-A-Week English with an Interspersed Elective," <u>California Journal of Educational Research</u>, 15:190-198, September 1964.

4. Heathers, Glen, "Research in Implementing and Evaluating Cooperative Teaching," <u>National Elementary Principal</u>, 44:27-33, January 1965.

5. Klausmeir, Herbert J.; Wiersma, William, "Team Teaching and Achievement," <u>Education</u>, 86:238-242, December 1965.

6. Lambert, Philip, <u>Classroom Interaction, Pupil Achievement, and Adjustment in Team Teaching as Compared with the Self-contained Classroom</u>, U. S. Department of Health, Education, and Welfare, Office of Education, Cooperative Research Project Number 1391, University of Wisconsin, Madison, 1964.

7. Norton, Daniel P., "The Potency of Educational Treatments," paper presented at the Presession of the Wisconsin Educational Research Association, December 1967.

8. Nystrand, Raphael O.; Bertolaet, Frederick, "Strategies for Allocating Human and Material Resources," <u>Review of Educational Research,</u> 37: 448-468, October 1967.

9. Zweibelson, I.; Bahnmuller, M; Lyman, L., "Team Teaching and Flexible Grouping in the Junior High School Social Studies," <u>The Journal of Experimental Education</u>, 34:20-32, Fall 1965.

EFFECT OF MASSIVE EDUCATIONAL

INTERVENTION ON ACHIEVEMENT

OF FIRST GRADE STUDENTS [1]

THOMAS M. GOOLSBY, JR.
University of Georgia

ROBERT B. FRARY
Regional Educational Laboratory for the Carolinas and Virginia

ABSTRACT

Two hundred Gulfport, Mississippi, first grade students received the benefit of extensive and intensive treatments designed to enhance educational effect. Approximately half of these students were from definitely deprived backgrounds and many demonstrated a very low degree of readiness for first grade activities. Comparison classes following the established first grade curriculum were also monitored to provide a basis for comparison. Results strongly suggest that the experimental treatments resulted in higher achievement. More important than the direction of the observed differences are their magnitudes. Enhancement of achievement due to the experimental treatments appeared greatest for low readiness students by a wide margin.

THE CLAIM has often been made that lack of adequate financial and administrative support for schools results in inferior achievement of pupils. The causal basis for such outcomes is easy to hypothesize but difficult to investigate. The reason for this difficulty is that only through the expenditure of large sums of money can the basic conditions be modified within a framework compatible with a strong research design.

There are studies too numerous to cite relating to the achievement levels attained by students in poorly financed school systems as compared to those in more adequately financed systems. In general, the conclusion seems to be that the better conditions, generally associated with better financing, result in higher achievement. This conclusion, plausible as it may seem, is not conclusive since no study comparing achievement in two settings can possibly account for all the differences in variables, other than those involving finances, which may affect achievement. Even if intelligence, income level of parents, and other accessible variables are held constant, the multitude of inaccessible variables may tip the scales in one direction or another. Factors such as community attitude toward schools, dietary differences, teacher training differences, etc., may have a strong cumulative effect.

The present study is unique in that a massive change in educational conditions was effected within a single community utilizing a research design which permits a statistically powerful evaluation of observed differences in achievement. While generalization from the findings may not be possible in all cases, a wide scope of generalizations is feasible. Specifically, the findings should be applicable wherever an urban school system faces the problem of lower achievement for segments of the system which have been subjected to inferior learning conditions.

PROCEDURES

The Gulfport, Mississippi, Municipal Separate School District received a United State Office of Education Title III grant of $135,882 for the study of approximately 450 first grade students during the school year, 1968-1969. The bulk of this money was to be spent on massive improvement of educational conditions, with the further requirement that half the students benefiting from improved conditions come from disadvantaged backgrounds.

To accomplish the above objectives, ten experimental classes were established in integrated and all-Negro schools. These classes were limited to twenty students; and each class was provided with an aid to handle routine tasks. The two hundred students

assigned to these classes were chosen randomly from the in-school populations, half from those known to be deprived and half from the remaining population. Teacher assignment, while not random in a formal sense, was made without any discernible source of bias. Classes, both homogeneous and heterogeneous with respect to readiness, were established based on total scores on the Metropolitan Readiness Tests (MRT), Form A (revised in 1964) (5). This measure had been obtained for all students of the district prior to assignment to classes.

Special experimental class materials designed to promote readiness and enhance the curriculum were developed by the University of Georgia Research and Development Center in Educational Stimulation for use in these classes. It was believed by teachers and administrators of the local system that the addition of these materials insured a more-than-adequate curriculum. The main features of the enhanced curriculum are repeated use of pretesting and posttesting to evaluate achievement, provision of immediate rewards, sequencing and structuring of learning activities, and supplementation of the traditional curriculum. Teachers and the community displayed interest and enthusiasm in the program.

Table 1 lists materials used in the experimental classes and those used in ten comparison classes which were chosen from the remaining twenty-four first grades of the district. These comparison classes contained twenty-eight to thirty-five students and received no special treatment during the year, except for testing which paralleled that for the experimental classes.

TABLE 1

MATERIALS USED WITH COMPARISON AND EXPERIMENTAL GROUPS

Materials Used with Experimental Groups*

A Beginning Geography Unit: Earth: Man's Home, Imperatore, W. (Publication No. 1BC).

Concept of Culture, Hunt, A.; Blackwood, J.; Emmons, F. (Publication No. 51a)

Language Arts and Verbal Learning Program: Part I: Introductory Exercises in Oral Language, Jennings, B. L.; Walter, P.; Duhling, D.; Quirk, K.

Language Arts and Verbal Learning Program: Part II: Introductory Exercises in Writing. Aaron, R. L.; Mason, G. E.

Language Arts and Verbal Learning Program: Part III: Introductory Exercises in Writing, Aaron, R. L.; Mason, G. E.

Mathematics Program: Suggested Mathematics Activities for 5-Year-Olds, Perrodin, A. F.

Music Program: Developing Basic Concepts of Music, Williford, B.; Simmons, G. M.
(Table 1 continued in next column.)

TABLE 1   (Continued from previous column)

Pre-primary Science Program, Level 1, Zeitler, W. R.

Physical Education Program: Movement Exploration, Gober, B.; Albertson, L.

Social Science Program: Getting Acquainted, Hunt, A.

The Ginn Basic Readers, Russell, D. H. and others, Ginn and Company, Boston, 1964.

Materials Used with Comparison Groups

Imaginary Line Handwriting, Steck-Vaughn.

Sets and Numbers, Gundlach, B. H.; Welch, R. C.; Buffie, E. G., Laidlaw Brothers, Atlanta, Georgia, 1965.

The Allyn and Bacon Basic Readers, Sheldon, William D. and others, Allyn and Bacon, Atlanta, Georgia, 1962.

The Ginn Basic Readers, Russell, D. H. and others, Ginn and Company, Boston, 1964.

This Is Music, Sur, William R. and others, Allyn and Bacon, Atlanta, Georgia, 1967.

The L. W. Singer Basic Reader, Pratt, Marjorie and others, L. W. Singer Company, Atlanta, Georgia, 1965.

The Scott Foresman Basic Readers, Gray, William S. and others, Scott, Foresman, and Company, Atlanta, Georgia, 1965.

Today's Basic Science.

We Live with Others, Hunnicutt, C. W.; Grambs, J. D., The L. W. Singer Company, Chicago, Illinois, 1963.

* All materials used with the experimental groups except the Ginn Basic Readers, were published by the Research and Development Center, University of Georgia, Athens, 1968.

It was possible to select or establish comparison classes sharing many of the characteristics of the experimental ones. However, administrative considerations precluded matching each group with respect to every relevant variable. For example, there were no integrated, heterogeneously grouped classes in the comparison set while there were three such classes in the experimental set. This design deficiency places limitations on the interpretation of certain outcomes of the study but by no means invalidates the conclusions generally. While the comparison classes cannot be said to have been randomly constituted, no effect on outcomes seems likely as a result of selection bias.

In addition to MRT scores, every effort was made to obtain data for every experimental and comparison subject for the following:

1. Ability Measure: Otis-Lennon Mental Ability Test, 1967, Primary II Level, Form J (7), administered December 1968.

2. Personal Data: (a) Number of Siblings; (b) Days Absent; (c) Ethnic Group (Negro or white); (d) Father's Occupation (professional, skilled, or unskilled).

3. Achievement Measures: (a) Metropolitan Achievement Tests, Primary I Battery, Form A (6) administered February 1969, Form B administered May 1969; (b) Botel Reading Inventory (2), administered April 1969; (c) Teachers Estimate of Reading Level. [2]

To analyze the achievement data, a multivariate analysis of covariance was performed. Mental age was the covariate and ten achievement scores for each student represented the criterion vector. The design was a complete factorial with four factors as follows:

1. Experimental Condition: Two levels, experimental and comparison.

2. Readiness: Three levels, according to total scores on the MRT—below 40 = Low Readiness; 41 to 65 = Medium Readiness; above 65 = High Readiness.

3. Father's Occupation: Three levels, professional, skilled, unskilled.

4. Sex: Two levels, male and female.

The multivariate analysis of covariance program from Multivariate Statistical Programs (3) was used for analysis. This program is very flexible in that unequal numbers of cases per cell are handled with appropriate adjustments for significance levels according to the method presented by Bock (1). The F-tests are based on Rao's approximation of Wilks' lambda criterion.

Because of certain departures from the formal requirements for the application of multivariate analyses of variance the reader is asked to make his own evaluation of certain outcomes. Nevertheless, these deficiencies are comparatively minor. For example, there is a slight tendency for the treatment to correlate with a covariate, mental age, due to the design requirement that half the experimental proportion come from deprived areas. This population is greater than the proportion of deprived in the comparison group. Also there was no random assignment between treatments although there was essentially random selection for each group. In spite of these shortcomings, the analysis presented is judged to be the most informative possible in view of the complexity of the data.

RESULTS AND DISCUSSION

In Table 2, it is seen that the main effect, experimental condition, accounted for differences in cell means at the .001 level of probability. An interpretation of this result may be made from inspection of Table 3.

It can be observed in Table 3 that for the covariate, mental age, the experimental group mean is 6.3 months less than that for the comparison group. In spite of this deficiency, the achievement means for the experimental group are almost as high as those for the comparison group for several achievement measures taken late in the school year. Further, after adjustment for the covariate, the experimental group's achievement means actually exceed those for the comparison group on eight of the ten achievement variables.

Table 2 also shows that the observed interaction between father's occupation and experimental conditions would occur by chance only 4.5 percent of the time. An inspection of the cell means shows that such an interaction is largely the result of the failure of children with fathers in the highest occupational category to achieve higher scores than those whose fathers are in the second occupational category in comparison classes after adjustment for the covariate. In experimental classes, children whose fathers were in the highest occupational category did achieve higher scores than those whose fathers were in the second occupational category after adjustment for the covariate. This condition was not prevalent with respect to all the achievement variables in the study. Only in the case of the first Metropolitan Achievement Test (MAT) Reading Subtest does the univariate F-test have an associated probability value of less than .05.

As shown in Table 2, the F ratio associated with the interaction between readiness and experimental condition can hardly be attributed to chance (p = .006). An inspection of the data in Table 4 readily accounts for this result. Note that the mean covariate scores (mental age) are nearly the same for the experimental and comparison groups under high readiness. Then, in the high readiness column, it is observed that differences between the experimental and comparison high readiness groups are negligible in most cases, especially for variables reflecting end-of-year achievement. Such is not the case for the low or medium readiness groups. In these two cases, the covariate favors the comparison group. Thus, substantially lower scores should be expected on most variables for the experimental group as compared to the comparison group. An inspection of the low and medium readiness columns shows some striking departures from this expected result. In fact, in many cases, the experimental group exceeds the comparison group. The significance of the interaction may be explained in this manner: Low and medium readiness groups profited extremely from the experimental treatment while high readiness students achieved approximately the same under either treatment. This result is extremely important in that it pinpoints the effectiveness of the experimental program. Moreover, it is a plausible result in that it has long been claimed that lack of attention has a much worse effect on less well prepared students than on those who can "cope for themselves."

Table 5 presents descriptive statistics for other variables in the study according to experimental and comparison groups. It was judged that the experimental group was considerably less well prepared for school. They came from homes where the father's occupation was of lower status, and families were large. They also had much weaker vocabularies. It is interesting to note, however, that there was no

TABLE 2

MULTIVARIATE ANALYSIS OF COVARIANCE USING WILKS' LAMBDA CRITERION AND RAO'S APPROXIMATION

Factor		df	F	p >
Main Effects:	Experimental Condition (EC)	(10, 391)	6.44	.001
	Father's Occupation (FO)	(20, 782)	2.80	.001
	Sex (S)	(10, 391)	2.48	.007
	Readiness (R)	(20, 782)	5.71	.001
First Order Interactions:*	EC x FO	(20, 782)	1.61	.045
	EC x S	(10, 391)	.951	.486
	EC x R	(20, 782)	2.00	.006
	R x FO	(40, 1485)	1.13	.266
	R x S	(20, 782)	.614	.905
	FO x S	(20, 782)	1.45	.091

* No higher order interactions significant.

TABLE 3

EFFECT OF EXPERIMENTAL CONDITION ON ACHIEVEMENT

Variable	Experimental N = 181 Raw Mean	SD	Comparison N = 264 Raw Mean	SD	Univariate F Tests F (1, 400)	p <
Mental Age, months (covariate)	73.2	13.4	79.5	11.8	—	—
MAT First Administration: Word Knowledge[a]	14.3	6.95	16.4	7.70	3.39	.066
Word Discrimination[b]	14.5	6.77	17.2	7.39	7.83	.005
Reading[b]	14.3	6.33	17.2	7.58	6.57	.011
Arithmetic[a]	33.3	16.0	38.5	15.0	1.35	.246
MAT Second Administration: Word Knowledge[a]	23.1	8.04	23.4	7.93	2.76	.097
Word Discrimination[a]	23.0	8.28	24.1	8.18	.131	.718
Reading[a]	24.4	10.5	24.8	10.8	4.36	.038
Arithmetic[a]	46.5	13.9	47.7	12.5	4.21	.041
Botel Instructional Level[a, c]	2.02	1.06	1.87	1.05	13.0	.001
Teacher Evaluation, grade placement	1.51	.293	1.58	.270	.865	.353

[a] Magnitude of adjusted means is greater for experimental group than comparison group. [b] Magnitude of adjusted means is greater for comparison group than experimental group. [c] Coded as follows: 1=preprimary; 2=primary; 3, first grade; 4= second grade, first semester; 5= second grade, second semester; etc.

TABLE 4

INTERACTION BETWEEN EXPERIMENTAL CONDITION AND READINESS BY LEVELS

Variable		Low Readiness MRT Total Less Than 41			Medium Readiness MRT Total 41-65			High Readiness MRT Total Above 65			Univariate F Tests	
		N	Raw Mean	SD	N	Raw Mean	SD	N	Raw Mean	SD	$F(2, 400)$	$p <$
Mental Age, months (covariate)	(E)[a]	91	65.8	11.1	44	76.4	7.81	43	86.7	10.2	--	--
	(C)[b]	71	68.8	10.5	96	80.4	8.91	92	87.0	9.10		
MAT First Administration:												
Word Knowledge	(E)		11.3	6.32		15.8	4.84		19.4	6.75	1.12	.326
	(C)		10.5	5.5		16.2	9.04		21.4	6.06		
Word Discrimination	(E)		10.9	4.94		16.2	5.70		20.4	6.36	1.60	.203
	(C)		11.7	4.96		16.2	6.59		22.3	5.86		
Reading	(E)		11.5	5.10		15.2	4.52		19.4	6.98	.988	.373
	(C)		14.5	6.75		17.6	7.72		18.6	7.76		
Arithmetic	(E)		23.3	12.9		39.6	9.07		48.7	11.3	.958	.385
	(C)		24.7	14.3		39.3	12.7		48.6	8.96		
MAT Second Administration:												
Word Knowledge	(E)		18.3	7.48		27.2	4.43		29.2	5.24	6.00	.003
	(C)		17.1	7.08		23.1	7.16		29.3	3.91		
Word Discrimination	(E)		18.2	7.73		26.5	5.53		29.5	5.13	1.97	.141
	(C)		17.0	7.14		24.3	7.87		29.8	4.45		
Reading	(E)		17.7	7.66		29.4	8.21		33.8	7.76	6.70	.001
	(C)		17.4	7.33		23.4	9.77		32.2	9.24		
Arithmetic	(E)		40.0	15.0		51.7	7.93		55.5	7.30	2.48	.085
	(C)		37.9	15.4		47.8	10.9		55.2	4.31		
Botel Instructional Level [c]	(E)		1.39	.628		2.41	.897		3.00	1.05	4.57	.011
	(C)		1.24	.520		1.73	.900		2.51	1.10		
Teacher Evaluation, grade placement	(E)		1.37	.235		1.56	.166		1.77	.306	.001	.999
	(C)		1.40	.190		1.61	.586		1.76	.264		

[a] E = Experimental   [b] C = Comparison   [c] Coded as follows: 1 = preprimary; 2 = primary; 3 = first grade; 4 = second grade, first semester; 5 = second grade, second semester; etc.

TABLE 5

MEANS AND STANDARD DEVIATIONS FOR OTHER VARIABLES BY EXPERIMENTAL CONDITION

Variable	Experimental N = 181		Comparison N = 264	
	Raw Mean	SD	Raw Mean	SD
Father's Occupation[a]	2.39	.679	1.94	.611
Number of Siblings	4.60	2.62	3.39	1.99
MRT - Total Score	42.4	20.2	51.0	18.5
Botel Potential Level[b]	2.83	1.88	4.24	2.58
Days Absent	10.0	10.8	10.9	10.9

[a] Coded as follows: 1 = professional; 2 = skilled; 3 = unskilled

[b] Coded as follows: 1 = preprimary; 2 = primary; 3 = first grade; 4 = second grade, first semester; 5 = second grade, second semester; etc.

substantial difference in number of days absent for the two groups.

It should be pointed out at this time that conclusions regarding the superiority of the experimental treatment are important mainly in the light of the magnitude of the changes effected. Literature is replete with studies which show that small class size is a factor in improving instruction. Clearly, however, class size alone can never be expected to account for large changes in educational outcome. Haberman and Larson (4) report that in the absence of other directions teachers given smaller than usual classes tend to present the same material in the same manner to their students as teachers who have larger classes. In the words of these investigators:

> Would cutting class size change instruction? We doubt it. Teachers just don't differentiate refinements of instructional activities; their role perceptions are probably not a function of class size at all. If smaller classes are to make a difference in the classroom behavior of teachers, it may be that they need to be instructed on how to teach a small class in different ways.

The Gulfport Project was designed not simply to show the effect of class size but to provide the teachers with a variety of material and activities to permit them to take advantage of reduced class size.

Again, of course, the literature is full of studies which show the superiority of one teaching method or one set of materials over another, but experience has shown that these results can seldom be replicated. With its combination of activities, a smaller class size and special materials, the Gulfport Project offers a model for producing substantial results in the improvement of instruction, without the problem of non-replicability caused by failure to account for major variables within the system. For example, how can class size be depended upon to improve instruction if there is no uniform provision of supplementary activities and material to permit the teacher to capitalize from the condition? Similarly, how can a new method or set of material be depended upon to influence instruction when teachers are forced to use the material under a variety of sometimes unsatisfactory conditions? The Gulfport Project avoids the pitfalls posed by these two questions.

From data not shown, the following additional outcomes were noted:[3]

1. Readiness by levels of high, medium, and low had a strong and highly significant effect on achievement. The high readiness groups were favored on every variable, even after adjustment for the covariate, mental age.

2. The lower readiness students have poorer scores on Father's Occupation, Number of Siblings, Botel Potential Level and Days Absent.

3. Father's occupation has a global effect on achievement. The computer printout shows that on most variables the skilled group was the best performer, taking the covariate, mental age, into consideration. This result strongly implies that the schools could obtain better results on the average for children from professional homes and quite possibly from unskilled homes as well.

4. Readiness grouping based on the MRT appeared not to have a beneficial effect on achievement of either high or low readiness students.

5. Negro students in integrated classes attained far higher achievement levels than those in all-Negro schools even after adjustment for the covariates, mental age and father's occupation.

FOOTNOTES

1. This research was pursuant to a USOE contract, a Title III ESEA Project Number 68-0671-0, in the Gulfport Municipal Separate School District, Gulfport, Mississippi, 1968-69.

2. Developed specifically for the Gulfport Project.

3. Readers who are interested in a more detailed reporting of these and other peripheral results should consult the technical report on the project by Thomas M. Goolsby, Jr. and Robert B. Frary, Enhancement of Educational Effect Through Intensive and Extensive Intervention - The Gulfport Project, Gulfport, Mississippi: The Gulfport Municipal Separate School District. Copies may be obtained by writing to Dr. Mercer Miller, Assistant Superintendent, Gulfport Municipal Separate School District, Gulfport, Mississippi.

REFERENCES

1. Bock, R. D., "Programming Univariate and Multivariate Analysis of Variance, Technometrics, 5:95-177, 1963.

2. Botel, M., Botel Reading Inventory, Follett Publishing Company, Chicago, Illinois, 1961.

3. Clyde, D. J.; Cramer, E. M.; Sherrin, R. J., Multivariate Statistical Programs, University of Miami, Coral Gables, Florida, 1966.

4. Haberman, M.; Larson, R. G., "Would Cutting Class Size Change Instruction," The National Elementary Principal, 47:(no. 4)18-19, 1968.

5. Hildreth, G. H. and others, Metropolitan Readiness Tests, Harcourt, Brace, and World, Inc., New York, 1964.

6. Metropolitan Achievement Tests, Harcourt, Brace, and World, Inc., New York, 1958.

7. Otis, A. S.; Lennon, R. T., Otis-Lennon Mental Ability Test, Harcourt, Brace, and World, Inc., New York, 1967.

EFFECTS OF ADJUNCT QUESTIONS, PRETESTING,

AND DEGREE OF STUDENT SUPERVISION

ON LEARNING FROM AN INSTRUCTIONAL TEXT

H. W. GUSTAFSON and DAVID L. TOOLE
Bell Telephone Laboratories, Incorporated

ABSTRACT

Fifty-eight students undertook self-paced study of a 23,000-word text. The 2x2x2 design involved: (1) adjunct questions versus no adjunct questions, (2) a pretest versus no pretest, and (3) supervised study versus unsupervised study. The posttest included the pretest items and adjunct questions along with new items. A retention test came 7 to 13 days later. Neither pretesting nor supervision had any influence on achievement. Adjunct questions did improve performance, but only with respect to that subtest of the posttest comprised of the adjunct questions themselves. Thus, adjunct questions failed to produce the general beneficial effects on learning observed elsewhere by others. This appears attributable to the experimental requirement of other studies that Ss read the instructional material only once.

THE EFFECTIVENESS of instructional materials depends not alone on properties inherent in the materials, but also on the circumstances surrounding their use. The amount learned from a given textbook, film, or other instructional exposure always can be influenced significantly by appropriate manipulations affecting either the learner, the exposure itself, or the environment in which the two interact. Beneficial applications of this principle are difficult, however, because so little is known about what constitutes appropriate manipulations. It is easy to depress learning—through exorbitant levels of ambient noise, for instance—but to enhance performance is another matter, particularly when the materials already are tolerably good and the environment benign.

The present experiment was concerned with assessing the learning consequences of possible manipulations relating to individual self-study of expository text. We were interested, though, only in manipulations that might practically be put to use in typical instructional settings such as the public schools, not with procedures demanding large investments of time, money, or talent. Available techniques of the former variety appear to fall in three general categories:

1. Precondition the learner (e.g., preview the material to be studied).

2. Modify the conditions of study (e.g., require note-taking).

3. Provide study aids (e.g., supply learning objectives).

The design, accordingly, was a 2x2x2 factorial[1] involving one variable from each category:

1. Pretest versus no pretest.

2. Scheduled, supervised study versus independent, unsupervised study.

3. Adjunct questions versus no adjunct questions.

It was felt (a) that pretesting might induce sets to attend to the material more assiduously, especially those portions directly related to the pretest questions; (b) that the discipline deriving from scheduled, supervised study might promote more

careful reading of the text; and (c) that adjunct questioning might lead not only to better learning of the material covered by the questions themselves, but to better study behavior generally. Adjunct questioning has long been advocated by Pressey (7) as a simple, inexpensive means of augmenting the instructional value of a text, but it has never been clear whether it is the text material that is learned, or only the adjunct questions. A number of recent studies (2, 3, 10) have demonstrated that if adjunct questions are employed in certain special ways, they can indeed produce general facilitive effects as well as the more predictable specific ones.

Pretesting was done as much to ensure experimental precision as to foster learning. Knowledge of results consequently was not supplied, nor were Ss told that the pretest was meant to serve as a learning aid.

METHOD

Subjects

From the senior class of Morristown High School, New Jersey,[2] fifty-eight volunteers were paid to serve as Ss. About half were female, half male. Each volunteer was paid a flat rate independent of the time it took to complete the required work. The rate, however, was made contingent on posttest performance. The sum paid was either $18, $15, or $12; the average time investment by the students, including orientation and testing, was about 7 hours.

The Text

The text studied was a 23,000-word nontechnical introduction to computers, prepared originally as the opening lesson of a self-study programming course for beginners. It provided a comprehensive overview of how one goes about using a modern computation center to solve scientific problems. No overt responses were required.

In addition to the textual material itself, the lesson included the following components:

1. A form for recording study time;

2. A detailed table of contents;

3. Fifteen technical illustrations (photographs, diagrams, examples of computer output, etc.);

4. A list of key terms and concepts, indexed to facilitate look-up;

5 A lesson summary, about one-third the length of the main text;

6. A moderately easy check-out quiz.

Previous investigation had established that the lesson was well accepted by students as a self-instructional instrument, and that the mean self-paced study time was around 5 hours (4.9 hours in the present instance). It was known also that the lesson was effective in accomplishing its instructional objectives.

Half the Ss studied the text in exactly the form

above; the other half studied it in conjunction with a separate booklet containing fifty-three adjunct questions, a mixture of true-false, multiple choice, short answer, and cloze items. In the adjunct-question treatment, the student encountered instructions after every two or three pages of the text directing him to answer either one or two of the questions in the booklet. The questions invariably came after the material to which they related, an arrangement regarded as important from the standpoint of securing general, as opposed to specific, learning effects (2, 3, 4, 5, 10). Knowledge of results was supplied by indexing each question to the page of the text where the answer could be found.

Examinations

Not counting the check-out quiz, which was not used as a criterion, three examinations were administered:

1. A 24-item pretest;

2. A posttest consisting of the twenty-four pretest items, plus the fifty-three adjunct questions, plus an additional forty-four new items (the forty-four new items were administered first).

3. A retention test of thirty-six items, all different from any of the foregoing.

The pretest and posttest were closed-book, power tests; the retention test was an open-book, timed test. All three examinations constituted representative samplings of the lesson content. To defeat guessing on the retention test, the latter was composed entirely of short-answer items. The pretest and posttest, however, each contained a variety of item types (like the adjunct questions).

Procedure

The main portion of the research was carried out, on school premises, during a 1-week spring recess; retention testing was done on a Saturday 1 week later. On Monday morning all Ss reported to the school auditorium and were randomly assigned to experimental treatments. A general orientation then ensued; ring binders were distributed containing the experimental materials; and half the Ss—those destined to study at school under supervision—were moved from the auditorium to the school cafeteria.

Both groups of students next were told to open their binders, remove "envelope A" and complete the materials found inside On opening envelope A, half the Ss found both the pretest and a background questionnaire; the remainder found only the "placebo" background questionnaire.

After finishing the foregoing task, the group engaging in supervised study were read instructions telling them they were to study in the cafeteria from 9:00 to 12:00 each morning until they had completed the check-out quiz and felt prepared to take the posttest. At that point they were to turn in their materials and receive an appointment for the posttest. In the meantime, they were not to be allowed to remove the instructional materials from the cafeteria.

The unsupervised group were being instructed

meanwhile that they could study the materials when and where they wished, and that a study room would be open at the school each day from 9:00 A. M. to 9:00 P. M. should they elect to avail themselves of it. They, too, were to turn in their materials and obtain a posttest appointment as soon as they felt prepared.

Thereupon, all Ss silently read additional instructions telling half of them simply to "go ahead and start studying," and the other half to open "envelope B" containing the booklet of adjunct questions together with directions for their use. The unsupervised group now were dismissed to do as they pleased, and the experiment proper got under way.

By the second day (Tuesday), a few of the students began to finish and were given appointments to be posttested Wednesday, approximately 24 hours later. This pattern of 24-hour posttest delay then continued through the week, with morning and afternoon testing sessions each day. Despite an entire week of glorious weather, the great majority of Ss were dependable and punctual in attending supervised study periods and keeping posttest appointments. In analyzing the performance of individuals posttested at different times during the week, no evidence of test compromise could be detected.

On Saturday morning, 1 week later, the students reassembled to take the retention test and collect their pay. Again most were punctual. Makeup sessions were arranged for absentees, the last of whom was finally given the retention test on the following Wednesday. The time between posttest and retention test varied from 7 to 13 days, the modal span being 9 days.

RESULTS

Pretest Results

Scores on the pretest (twenty-nine Ss) were approximately at the level of blind guessing. A 2x2 analysis of variance showed no significant interaction or main effects.

Readministration of the pretest items as part of the posttest yielded a test-retest correlation of -0.08 and a mean gain more than eight times its standard error. Together these findings suggest that substantial learning took place as a consequence of study.

Posttest Results

The experiment began with sixty-four Ss, six of whom dropped out. Unfortunately from a statistical point of view, three of the dropouts occurred in the same cell of the 8-cell design. We attribute no significance to this (if all distributions are equally probable, there is better than a 20 percent chance that some one of eight cells will contain three or more of six dropouts), but it did result in a fairly acute case of unequal N's. Five analyses of variance were therefore conducted, one using 0's for the missing scores, another using cell means, and three involving different random discards of data so as to bring all cells down to N = 5.

The three parts of the posttest—pretest items, adjunct items, and new items—were analyzed sepa-rately. Means and standard errors for the treatment main effects are shown in Table 1.

Scores on both the pretest items and new items were found to exhibit a tendency toward either triple interaction or double interaction between the pretest/ no-pretest and supervision/no-supervision factors. We do not feel, however, that this tendency was of sufficient strength to invalidate straightforward interpretation of the main effects. (For example, of ten F-ratios for the double interaction—one for pretest items and one for new items in each of five analyses—only one reached the .05 level of significance.)

All main effects for the pretest items and new items proved nonsignificant, the F-ratios in all analyses being uniformly < 1. For these two parts of the posttest, thus, the null hypothesis appeared fully satisfied. Furthermore, the pretest items and new items behaved very similarly from a statistical standpoint throughout all analyses; it was as if the pretest items in the posttest were simply another set of new questions just as unfamiliar to the pretested group as to those who had never seen them before. Additional evidence of the similarity between the pretest items and new items is afforded by the correlation of 0.86 obtained between them.

The adjunct items in the posttest indicated no triple interaction but did display a tendency toward double interaction the same as for pretest items and new items. Here again, however, we adjudge the trend too weak to interfere with normal interpretation of main effects.

In four of the five analyses the adjunct items, not surprisingly, generated a highly reliable main effect in favor of the group who used the adjunct questions while studying the text (smallest F = 10.2, 1/32 df, $P \approx 0.006$). The only analysis in which this effect failed to show up strongly was the one that used 0's as the scores of the six missing Ss, but even there it was significant at about the 0.09 level (F = 3.1, 1/56 df).

Two of the analyses indicated a significant difference on the adjunct items between the pretested and not-pretested groups. However, one of these is suspect because it is the analysis that used cell means for the missing scores (F = 6.1, 1/56 df, $P \approx 0.02$), and the other, a "random discard" analysis, showed the difference only barely reliable (F = 4.7, 1/32 df, $P \approx 0.04$). Thus, in light of its statistical frailty, and equally because it seems to make no intuitive sense, we believe this finding ought to be disregarded.

Retention Test Results

The retention test (see Table 1) resulted in total satisfaction of the null hypothesis for all interactions and main effects. If the interaction tendencies observed at the time of the posttest were genuine, they evidently disappeared during the 1- to 2-week retention interval.

Test Reliability

Kuder-Richardson Formula 20 yielded coefficients from 0.87 to 0.93 for the three parts of the posttest. Adequate reliability for the retention test is indicated by its total-score correlations in the mid-to-high 70's

TABLE 1

MEANS AND STANDARD ERRORS BY TREATMENT MAIN EFFECT AND PERFORMANCE MEASURE

Main Effect	Performance Measure				
	Pretest Items (24 max.)	Posttest Adjunct Items (53 max.)	New Items (44 max.)	Retention Test (36 max.)	Study Time (in hours)
Pretest	14.3	37.4	27.6	20.0	4.98
(N=29)	(±1.0)	(±2.0)	(±1.4)	(±1.3)	(±0.43)
No Pretest	15.4	42.4	28.3	22.3	4.77
(N=29)	(±0.9)	(±1.3)	(±1.2)	(±1.1)	(±0.23)
Supervision	14.9	40.5	28.1	20.3	4.84
(N=31)	(±0.8)	(±1.4)	(±1.1)	(±1.0)	(±0.20)
No Supervision	14.7	39.2	27.8	22.1	4.92
(N=27)	(±1.1)	(±2.1)	(±1.5)	(±1.4)	(±0.47)
Adjunct Questions	15.4	44.8	28.6	22.2	4.58
(N=28)	(±0.9)	(±1.4)	(±1.2)	(±1.2)	(±0.25)
No Adjunct Questions	14.3	35.3	27.3	20.2	5.15
(N=30)	(±1.0)	(±1.6)	(±1.4)	(±1.1)	(±0.40)

with these various parts of the post-test.

Study Times

The time data produced a number of significant interactions—particularly a triple interaction—which are difficult to interpret. The group, for example, that was supervised but was not pretested and used no adjunct questions consumed nearly 4 hours more time on the average than its "opposite," the group that was not supervised but was pretested and did use adjunct questions (6.7 hours as against 2.9). Looking at the marginals, however, no substantial time advantage accrued to either level of any main effect (Table 1). The largest overall time difference was that for the adjunct-question factor, where the Ss using adjunct questions unexpectedly completed their study some 34 minutes faster than the group not using adjunct questions.

DISCUSSION

Effects of Pretesting

Whereas a pretest might be expected to enhance motivation, induce learning of the items in anticipation of the posttest, and/or focus the learner's attention on specific parts of the material to be studied, all achievement measures indicated that pretesting

was ineffectual in these respects. Assuming that pretesting does have beneficial effects sometimes, perhaps the circumstances that negated them in this instance were the length and intricacy of the study task (i.e., any potential effects may simply have extinguished with time) and the students' initial low level of knowledge of the subject matter (i.e., the pretest questions may have appeared meaningless).

Effects of Supervision

Supervised study, as opposed to independent study, was likewise ineffectual. The only real hint of any difference between the two treatments was that five of the six dropouts occurred in the unsupervised group. This is not, however, a statistically significant difference (p > .20).

It is noteworthy that the study facility provided at school for the unsupervised group was almost never used. Precise records were not kept, but it is known that fewer than a dozen students ever visited the room, none of them occupying it for more than an hour or so in all. There seems to be a clear implication that high schools might be able to take more advantage of their students' capacity for independent self-study than they presently do. The data are consistent with the growing contention (e.g., 8) that a large proportion of the secondary curriculum could be acquired satisfactorily through "learning without teaching."

Effects of Adjunct Questions

The sole clear-cut finding of the study was that students exposed to adjunct questions achieved significantly better than those not exposed, but only with respect to that subtest of the posttest comprised of the adjunct questions themselves. Adjunct questioning did not affect performance on any other part of the posttest or on the retention test. It thus appears that the adjunct questions, though generating strong question-specific effects, failed to produce the more general beneficial effects on learning observed in recent experiments by others.

As Rothkopf (9) has pointed out, however, adjunct questions may have either beneficial or detrimental general effects. If both types of effect were operative in the present case, one might reasonably expect to find some cluster of new items in the posttest on which the adjunct-question group displayed superior performance, and another cluster on which their performance was inferior. To check this out, each of the forty-four new items of the posttest was analyzed, with the aid of the Lawshe nomograph (6), to determine if the proportion of Ss passing the item in the adjunct-question group was significantly larger or smaller than the proportion passing it in the no-adjunct-question group. It turned out that the two proportions differed at or beyond the .05 level on exactly four of the forty-four items, three times in favor of the adjunct-question group, once in favor of the other group. Since these results obviously are well within chance limits, this analysis further supports the view that adjunct questioning produced no general effects.

General effects typically have been elicited by adjunct questions in other investigations only when the questions are placed—as was done in the present study—after the text material to which they pertain. In all these other experiments, moreover, the Ss have been allowed to read the instructional material only once. Thus, the learner realizes he will have but a single chance at the material and is either told or soon becomes aware that questions of unknown content await him from time to time. It is understandable that these conditions should serve to heighten his concentration.

But when he knows he can reread the text and review the questions as often as he likes—which, after all, is the normal process in preparing for an examination—there is no longer a good reason for the student to concentrate any harder on the text than he would if not aided by adjunct questions. Hence, we believe, the adjunct questions lose their potency except as emphasizers of the specific information with which they deal. By the same token, it seems a likely surmise that the placement of the questions also loses its importance under "real life" study conditions. A guess is that the adjunct questions would have had the same impact in the present experiment no matter where they occurred in relation to the text.

CONCLUSIONS

Pretesting done without feedback on unfamiliar material proved in this study a poor way to precondition the learner to the task ahead. Had feedback been provided, and had the students been instructed to pay close attention to the questions even when they did not understand them, the outcome might have been more favorable. There is a distinct possibility, however, that any pretest containing specific, detailed questions is of little conceptual help to a learner. A better approach might be a highly general preview, or "advance organizer," of the type investigated by Ausubel (1).

The study found no evident advantage for regularly scheduled, supervised study at school over unscheduled, unsupervised study done mostly at home. Hawthorne effects undoubtedly played a role in sustaining the performance of the students, but any such effects, we feel, were heavily counterbalanced by the magnificent weather enjoyed throughout the period, a vacation week, in which the study was carried out. Our conclusion is that the rigorous attendance rules and lock-step scheduling so commonly invoked by high schools may be selling many students short.

On the strength of the present findings, we are obliged to conclude that adjunct auto-instruction offers little promise except as a device to foster learning of the adjunct questions themselves. The transfer induced by adjunct questioning was insufficient to improve performance even on the many pretest items, new items, and retention-test items that were related in content to various of the adjunct questions. All the same, the specific learning produced by the adjunct questions was not detrimental to achievement on these other tests and came at no extra cost in study time. Hence, an affirmative recommendation for the judicious use of adjunct questions seems indicated.

FOOTNOTES

1. More accurately, the study was an independent sub-experiment of a larger 2x2x3 factorial. The third level of the last factor is omitted from the discussion because it involved a complicated re-configuration of the text which is cumbersome to explain and which, in any event, proved ineffectual.

2. We thank the officials of the Morristown schools for their cooperation and assistance, notably, Dr. Harry W. Wenner, Superintendent of Schools, and Mr. William E. Kogen, Principal of Morristown High School.

REFERENCES

1. Ausubel, D. P., "The Use of Advance Organizers in the Learning and Retention of Meaningful Verbal Material," Journal of Educational Psychology, 51:267-272, 1960.

2. Bruning, R. H., "Effects of Review and Test-like Events Within the Learning of Prose

Materials," Journal of Educational Psychology, 59:16-19, 1968.

3. Frase, L. T., "Learning From Prose Material: Length of Passage, Knowledge of Results, and Position of Questions," Journal of Educational Psychology, 58:266-272, 1967.

4. Frase, L. T., "Effect of Question Location, Pacing, and Mode Upon Retention of Prose Material," Journal of Educational Psychology, 59: 244-249, 1968.

5. Frase, L. T., "Some Data Concerning the Mathemagenic Hypothesis," American Educational Research Journal, 5:181-189, 1968.

6. Lawshe, C. H.; Becker, P. C., "Three Aids in the Evaluation of the Significance of the Difference Between Percentages," Educational and

Psychological Measurement, 10:263-270, 1950.

7. Pressey, S. L., "A Puncture of the Huge 'Programming' Boom?" Teachers College Record, 65:413-418, 1964.

8. Pressey, S. L., "Two Basic Neglected Psycho-educational Problems," American Psychologist, 20:391-395, 1965.

9. Rothkopf, E. Z., "Some Conjectures About Inspection Behavior in Learning from Written Sentences and the Response Mode Problem in Programmed Self-instruction," Journal of Programmed Instruction, 2:(no. 4)31-45, 1963.

10. Rothkopf, E. Z., "Learning from Written Instructive Materials: An Exploration of the Control of Inspection Behavior by Test-like Events," American Educational Research Journal, 3:241-249, 1966.

SOME CORRELATIONAL ASPECTS OF PERFORMANCE ON THE ART SCALE OF THE WFPT AMONG CERTAIN VARIABLES IN A DEAF POPULATION

GERALD JOHNSON
Pennsylvania School For The Deaf

WILLIAM BRADLEY
The Pennsylvania State University

ABSTRACT

Because of low verbal skills exhibited by deaf subjects and because of earlier research results linking verbal behavior to aesthetic behavior, this preliminary investigation sought to lay some groundwork for an empirical study of the relationship between these two ordering systems. This study states the postulates upon which our working paradigm is based and presents the results of a correlational analysis of syntactic and aesthetic preference scores of ninety-two deaf high school students (both profoundly deaf and legally deaf). The results appear to corroborate earlier findings linking syntactic and aesthetic behavior. The study suggests the direction for an experimental study which is now in operation under a university grant.

THE PARADIGM upon which this correlational study was based is related to the way in which humans appear to organize syntax, metaphors, analogies, and other inarticulate systems. Based on previous research by the authors, there are some indicators which seem to suggest a complicated interconnection between these various ordering functions of the mind. The process of verbalization—that is the search for word-phrase counterparts to experience—appears to be related to the inarticulate ordering systems as an assimilating activity which seeks out the appropriate order and intent from among the possibilities available. The study of this process during the pre-committal stages of ordering may provide a fruitful path into the study of the systems themselves and their illusive interconnections. In art, it is as though the act of manipulating preconscious alternatives increases the flexibility of the S permitting him to perceive greater figural possibilities in the more ambiguous drawings. Just what this relationship is defies current analysis, but the data seem to indicate that both variables—verbal activity and figure preference—are dependent upon each other. It is known that the process of verbalization does indeed influence aesthetic behavior both in the act of art itself and in the S's preference for figural stimuli. But whether or not the preconscious strategies which come into play before a syntactic commitment is made are similar to the preconscious strategies involved in aesthetic ordering is not known. Nor is it known whether the stimulating of such preconscious strategies in one ordering system has a concomitant influence on the other systems.

Our working model suggests that such a concomitance might exist and that when Ss are permitted to verbalize about their own strategies they literally expand their time of pre-committal manipulation of alternatives. It is just this flexibility which may affect the quality of all inarticulate products. The indicators received from two previous studies suggest that this is apparently the sequence. In one study with hearing Ss there was a strong indication that such intrinsic verbal feedback has a positive influence on art production quality (3). A second experimental study suggested a similar relationship between the verbal feedback condition and figure preference although personality factors were involved in changes reported in the second study (2).

This paradigm is related to two basic observations about the nonverbal expression of experiences.

First, in the essentially less-verbal world of the deaf rests, perhaps, some of the illusive answers to the "how" questions in learning. For if learning is to take place at all among deaf Ss, it must be pedagogically correct. The subtle and unknown methods by which hearing Ss learn can slip by the most conscientious researcher while he incorrectly assumes that what he is controlling in his experiments is responsible for the observed change. But the margin of error appears to be substantially reduced for the study of the less-verbal learner. If classroom pedagogy is inadequate in the hearing classroom, learning still seems to take place. But, if classroom pedagogy is inadequate in the nonverbal classroom, learning simply does not occur at an appreciable rate.

Second, the apparent relationship of the inarticulate nature of deaf learning to the inarticulate ordering function of the plastic arts, suggests that once we know the extent of this interaction, we may be in a position to discuss the larger question of how aesthetic ordering relates to word meaning, concepts, syntax development, and reading achievement. The benefits of such information appear obvious.

Many ideas and concepts appear under the rubric of aesthetic experience requiring perhaps a conceptual definition for the reader before discussing the implications of the analysis herein reported.

In all definitions of aesthetic experience, one undercurrent seems agreeable. It apparently serves as a form of inarticulate communication requiring at least two of three parts: (1) the sensed or the initial reaction of the artist to a stimulus, (2) the object or art work which derives as a tangible expression of the sensed or experienced, and (3) the observer who relates this new object to his own wealth of experience which may result in a pleasurable or an unpleasant sensation.

The aesthetic experience differs from other communicative acts in that it primarily deals with nonverbal aspects, not unlike facial expression, gesticulation, and body movements.

It might be stated that any experience enters the mind at some irrational level which may immediately vent itself in some physical sensation or be transformed into something more rational. It must somehow be related or forced into a shape which will be compatible with the life model of the receiver. In such a transformation, distortion is likely to occur. It is precisely this distortion—taking on a physical shape and substance—which evolves into a form metaphor, or an image-concept of the experience. Should the S continue to rationalize his experience, he becomes logical about it, trimming it as much as possible by eliminating any semblance of its original irrationality. This distorted, yet compatible, form metaphor has great potential for communication at the nonverbal level. In other words, it is inarticulate communication, deriving its strength from common sources of feeling.

The salient point is that the aesthetic ordering function appears to be related to one of the earliest and perhaps most pregnant stages of the intellection process. It is here that the greatest flexibility in the interpretation of experience can occur. Flexibility has been regularly cited as a quality or at least a primary influence on creative behavior (6,8). One of the distinctions between the artist and the non-artist isolated by empirical scales has been this characteristic (9).

The work of Anton Ehrenzweig (4) related to the development of articulation suggests that the automatic assembly of visual and aural impulses from direct experience can be—indeed in art schools is—taught by eliminating the gestalt figural tendency through conscious effort toward flexible behavior. The art student is trained to reconsider—all at the inarticulate level—the numerous possibilities for his particular form metaphor.

After a point of committal is reached the mind appears to enter a refining phase. The form metaphor may take a plastic shape as it is or be refined through various stages until it has been condensed into logic. One method by which this refinement is accomplished appears to be associated with verbal behavior. Once again the empirical cues suggest that the most verbal students or at least those who score higher on verbal scales of intelligence also see more possibilities in partially articulated drawings or drawings which are not clearly defined or regular. In addition to this interconnection, it is becoming increasingly clear that planned verbal involvement in the art classroom has a positive influence on aesthetic growth (1,2,3,5). Indeed some language researchers suspect that the word itself, (both inarticulate and articulate) as a symbolic device, enables humans to categorize and synthesize real experience by permitting an abstract bridge between the rational and the irrational. The word as a symbolic function may be a closer relative to the aesthetic ordering processes than has been previously supposed. Luria suggested that the word did indeed serve such a function over and above its more obvious role of indicating objects (7).

With this developing paradigm coupled with carefully calculated experimental designs, it is hoped that some patterns of behavior will emerge which can be systematically controlled by classroom pedagogy.

This correlational study was designed simply as an exploratory analysis which would corroborate or discredit findings relating verbal behavior to figure preference. The deaf population of the upper school population of the Pennsylvania School for the Deaf provided a population where variation in verbal skills was minimal. Thus, any relationship detected in this group would be a more powerful indicator than in previous experiments where verbal variability was great. Table 1 describes the results of a correlational study which was intended to identify relationships between the following variables: age, sex, figure preference, IQ, degree of deafness, and three scores on the Stanford Achievement Test (SAT). The P-values reported represent only those which reached a significance level of .95 or above for the population involved.

FINDINGS OF STATISTICAL SIGNIFICANCE

The IQ factor correlated strongly with all the achievement scales. Since all of these scales were developed to reflect IQ such a correlation result was predictable.

Age represented a significant variable when

TABLE 1

CORRELATION MATRIX BMDO2D CORRELATION WITH TRANSGENERATION*

	Age	Sex	WFPT	IQ	DD	SAWM	SAPM	SAAC
Age	—							
Sex	—	—						
Welsh Figure Preference Test - Art Scale (WFPT)	.2409	-.2102*	—					
Intelligence Quotient (IQ)	—	—	—	—				
Degree of Deafness (DD)	—	—	—	—	—			
SAT Word Meaning (SAWM)	—	-.2353	.3533	.3257	—	—		
SAT Paragraph Meaning (SAPM)	.1955	-.2548	.3120	.3390	—	.8436	—	
SAT Arithmetic Computation (SAAC)	—	—	.2721	.4148	—	.6827	.7164	—

*Negative percentages reported on the sex variable refer to females.   N = 92 (Ages 14-18).

P-value Interpretation

P (r > .173) = .95
P (r > .242) = .99
P (r > .338) = .999

correlated with the measure of figure preference and paragraph meaning on the achievement test. It must be remembered, however, that the population being discussed is deaf. In normal hearing populations the WFPT art scale is considered to be free of the age factor. It was unusual to find a significant age factor for this age group related to paragraph meaning among the deaf. It had been assumed that the achievement test in use did not reflect this factor.

The sex factor was predictable in the hearing population and perhaps was corroborated in the deaf population. Females did significantly better on the WFPT art scale and on both word meaning and paragraph meaning although IQ and degree of deafness were not significantly correlated with the sex factor.

SIGNIFICANT CORRELATIONS WITH THE BARRON ART SCALE

In the deaf population a correlation between age and preference for the complex-asymmetrical figures was indicated. This relationship was unexpected as it had been assumed that the deaf population would be similar to a hearing population in this regard.

As expected, sex was correlated with the figure preference score. Females tend to score higher than males.

Significance at the .001 level was indicated when the Art Scale was correlated with the SAT word meaning achievement score. The Paragraph Meaning score on the achievement test was significantly correlated with the Art score at the .01 level. These correlations tend to support the research findings in

normal hearing populations related to verbal involvement and level of preference for complex-asymmetrical figures. They suggest a similar relationship between reading skills, verbal fluency, and high scores on the WFPT Art Scale.

The figure preference variable seemed to correlate with the other ordering functions as measured by word meaning, paragraph meaning, and arithmetic computation skills on the achievement test. Therefore, those deaf students who are achievers in these areas tend to score higher on a test of aesthetic preference.

The interpretation of these results is limited by two factors. First, the rho's reported in the analysis are not high—although they are significant for the population size—and, second, the correlations from this study do not imply cause-effect relationships but merely show the variables to be in some way related. Therefore, we will limit our interpretation to a more conceptual framework rather than try to establish a behavioristic paradigm.

As mentioned in the introduction of this article and the subsequent reference to our working model or paradigm, it seems clear that the assumptions made concerning the use of intrinsic verbal feedback might hold for non-hearing as well as hearing Ss. This by itself, holds a small promise for the study of specific verbal behaviors. It is also evident that syntactic development has either an indirect or direct relationship with aesthetic behavior—there is a connection.

From a research standpoint, both of these observations are important. We should now be free to

tackle the monumental problem of isolating various aspects of verbal acuity as it relates to perceptual or at least aesthetic sensitivity and researchers should be, in some measure, ready to mount an assault on the relationships which must surely exist between the ordering systems mentioned.

Pedagogically, one can look forward to the isolation of stimulus-response entities related to the use of art, theatre, and creative writing, etc., for the development of word-paragraph concepts and syntactic sharpening.

An experimental study is currently being planned by the authors to determine the effect of manipulating verbal behavior during the art process, on change in complex figure preference and achievement scores.

Our current prediction equation suggests that the stimulating of both articulate and inarticulate verbal activity by involving the deaf student in his own evaluative feedback in art will result in a positive change on our criterion measures.

REFERENCES

1. Beittel, Kenneth R., Effect of Self-Reflective Training in Art on the Capacity for Creative Action, Cooperative Research Project Number 1874, U.S. Office of Education, 1964.

2. Bradley, William, "An Experimental Study of the Effect of Three Evaluation Techniques and Personal Orientation on Aesthetic Preference," unpublished PhD thesis, University of Minnesota, Minneapolis, 1968.

3. Bradley, William, "A Preliminary Study on the Effect of Verbalization and Personality Orientation on Art Quality," Studies in Art Education, National Art Education Association, 9:(no. 2) 31-38, 1968.

4. Ehrenzweig, Anton, The Psycho-Analysis of Artistic Vision and Hearing, George Braziller, New York, 1965.

5. Getzels, J. and Csikszentmihalyi, M., "Creativity in Art Students: Personality, Cognition, and the Process of Discovery," a preliminary report read at the Annual Meeting of the American Psychological Association, Chicago, 1965, pp. 1-6.

6. Guilford, J.P. and others, "A Factor Analytic Study of Creative Thinking II: Administration of Tests and Analysis of Results," Reports from the Psychological Laboratory, Number 8, University of Southern California, Los Angeles, 1952.

7. Luria, A.L., "The Directive Functioning of Speech in Development and Dissolution," Word, 15:341-352, 1959.

8. Torrance, E.P., "Explorations in Creative Thinking," in Taylor, C.W. (Ed.), The Third University of Utah Research Conference on the Identification of Creative Scientific Talent, University of Utah Press, Salt Lake City, 1959.

9. Welsh Figure Preference Test Research Edition, Consulting Psychologists Press, Palo Alto, California, p. 13.

BOOK REVIEWS

continuum may not be totally different in their theoretical basis." Too many teachers, counselors, therapists, and school psychologists think just this, unfortunately, and do not recognize that there are clear-cut differences that do not permit a person to espouse both theoretical positions. Either one believes that there are internal mechanisms or intervening variables such as the id, ego, and superego, or one does not.

The book is valuable in that it attempts to delineate what has been experimentally validated and to differentiate this from the areas in which opinion is the mainstay. As Woody says, "published opinions of professionals do not constitute experimental research...." Too frequently, it seems, medical opinions and not the results of experimentation have been used as diagnostic criteria, and this is an unscientific practice. It is to be hoped that Woody's book will encourage those in psychology and education to follow where the experimental research leads, rather than continue to follow their own favorite theories, despite evidence to the contrary.

Professor Julia R. Vane, Reviewer
Director of Doctoral Programs
Hofstra University
Hempstead, Long Island, New York

LEARNING EFFICIENCY

OF STUDENTS IN VARYING ENVIRONMENTS[1]

JOHN A. LUCAS
The University of Tennessee

ABSTRACT

In determining whether or not students learned more effectively in different learning environments, experimental and control classes were conducted over a 2-year period in a basic statistics course that involved ten faculty members and 1,172 students. Grades, common examinations, and a checklist were the basic criterion measures used; mathematics background and overall grade point (GPA) served as control mechanisms. Major conclusions were: students preferring some type of independent study consistently underachieved; instructor types, which produced high achievement levels for specific types of students, were identified in behavioral terms; it probably will take drastic manipulations of the learning environment beyond the normal constraints of the university in order to produce effective changes in the learning pattern.

IN THE past the nature of the learning process has been studied by concentrating on such factors as feedback, pace, spacing, degree of immersion, degree of participation, mode of presentation, r e i nforcement, sequence, etc. Interest has centered on the external control mechanisms of the learning environment. Research in this manner, has, however, produced three areas of concern: first, in many cases characteristic differences in instructors have produced greater learning effects than were produced by manipulating the external factors in the learning environment; second, individual differences in learners have rarely been considered; third, seldom has the learner been consulted on how he would like to learn a given subject matter.

In this study, the focus was on identifying instructor characteristics that differentially effect learning. Learner characteristics were also explored to see how they interact with the different learning environments. Learners were interviewed as to how they might best learn. Thus an attempt was made to determine if these preferences are important to the learning process.

Since 1960, McKeachie (9) and Katz and Sanford (8) have stressed the need for considering student characteristics in determining optimum teaching methods; much of the research since then has reflected this thinking. Siegal (15) stated that one of the major deficiencies in research of this kind is the failure to consider different learning environments for different types of learners.

Doty (3) measured five student characteristics, including creativity, and identified a type of student who learned optimally under each of three instructional methods. In a study comparing programmed learning to a conventional method, Erickson (5) determined how measures of tolerance to ambiguity, rigidity, and some of the Minnesota Multiphasic Personality Inventory scales interacted. Using the 16-PF, McKeachie (10) found ways in which personality traits of students affected their learning under different instructors who were differentiated by observational techniques. Initial state of knowledge was shown by Shuford (14) to interact with f i v e instructional strategies.

As of this writing, some major research projects now underway indicate the importance of the interaction of instruction method with student characteristics. The University of Colorado (2) is conducting a project

which is intended to measure the effect of personality on the usefulness of programmed learning. A project at Purdue (13) is determining the effect of different types of programmed instruction on various subgroups of students differentiated by personality and ability profiles. The University of Oregon (6) is measuring the interaction of four instructional formats with thirty different personality, motivational, and attitude inventories in maximizing course achievement.

PROCEDURE

This project at The University of Tennessee ran for 2 years, encompassing 1,172 students and ten instructors. The pilot phase consisted of interviews and open-ended questionnaires which sought to determine how students wanted to learn. During the experimental phase the following data was gathered: (1) student academic background (grades and courses taken), (2) age of students, (3) academic field of study, (4) student preference for type of learning environment, and (5) score on common exam and grade in statistics course. An achievement score was determined by subtracting a predicted grade or exam score from the actual score. The predicted score was calculated using a multiple regression equation which included cumulative GPA, math GPA and a quality math background index (number of advanced math courses taken) as the predictor variables. The correlations between predicted and actual scores ranged from .55 to .60.

It was thought that grade or exam achievement was not a sufficient criterion measure and that some type of student rating of learning effectiveness was needed. Anderson (1) concluded that student rating and achievement each measured something different. Jones (7) also agrees student ratings and achievement are independent measures. Similarly, Remmers (11) found no relationship between opinion of the teacher and rank in class.

However, even given that a student's rating of his teacher is an independent criterion measure, there are many types of rating scales that could be used. Most rating scales use some type of overall response to the teacher. Whitlock (17) argues for the use of performance specimens as the basis of teacher performance evaluation. He feels that performance specimens have the advantage of being directly observable, having built in relevance, being related by a power function to the response criteria, and identifying specific areas for remedial action. Douglass (4) used such a performance specimen checklist and found that there was a positive relationship between measures of teacher effectiveness which resulted from the checklist and student achievement. However, the relationship was not so strong as to deny checklist measures as being independent of achievement.

During the final quarter of this project, the Teacher Performance Checklist, developed by Douglass (4) was used. This checklist contained seventy-seven behavioral performance specimens. The student was asked to mark only items which he personally observed in the course. Furthermore, the student was asked to assign a positive number if he thought the behavior contributed to the learning process, a negative number if it detracted, and a zero if no affective feeling was produced. For each person these weights

were added together to produce a summation score. The mean stimuli intensity score was derived by dividing the summation score by the number of items marked. At the end of the checklist, the student was required to make a global rating of the overall effectiveness of the teacher.

Four different experimental sections were used in the basic statistics course at different points in the project. Two of these sections were oriented toward the theoretical. The emphasis in these sections was on derivations and understanding concepts rather than on problem solving. Another experimental section promoted more independent study by meeting only 1 hour per week instead of 3. Students were given extensive course outlines so they could do much of the work on their own. The fourth experimental section was designed specifically for personnel management majors. Many of the illustrations and problems used in this section were taken from the personnel management field.

The three major objectives of the project were: (1) determine if there was an interaction between student characteristics and learning environment with regard to learning efficiency, (2) describe different learning environments in behavioral terms by use of the Teacher Performance Checklist, and (3) analyze the relationship between the various criterion measures used.

RESULTS

Several student characteristics that helped describe how students learned under different environments were identified. One was a student's preference for a learning environment. Approximately one-third of all the students entering the basic statistics course preferred some type of independent study while the remaining student body opted for the regular 3-day-a-week class. Students preferring some type of independent study achieved at lower levels than regular preference students in twelve of the fourteen sections offered over a period of 1½ years (see Table 1). These sections included eight different instructors and several experimental situations. Since this would happen by chance less than two in a hundred times, it might be inferred that students preferring some type of independent study are poorer achievers under a variety of learning environments and that it is difficult to develop an environment under which this preference group can perform at its maximum.

Another student attribute that appeared to help clarify learning patterns was age. Table 2 indicates that older students achieved at higher levels than younger students under instructors A, B, and C. This achievement difference is significant at the .02 level. There is little difference in achievement level between age groups for instructors D and E. Older students also rated instructors A, B, and C higher on the Teacher Performance Checklist than did younger students (see Table 3). Both differences between the two age groups rating instructors A, B, and C on the summation and mean stimuli intensity scores are significant beyond the .05 level. There is little difference in the ratings of instructors D and E by the two age groups. There then seems to be established two distinct groups of instructors. One produces a learning environment which is better for older students while the other performs equally well for both young and old.

TABLE 1

STUDENT PREFERENCE VERSUS INSTRUCTOR IN REGULAR CLASSROOM SITUATIONS ACHIEVEMENT DIFFERENCES

Instructor	Date	Criterion Measure	Students Preferring A Regular Class		Students Preferring Partial Independent Study	
			N	Mean	N	Mean
A	Winter 1969	Common Exam Achievement	51	+2.20	32	+ .13
A	Winter 1968[a]	Common Exam Achievement	13	+ .29	20	+9.62
A	Winter 1968[b]	Common Exam Achievement	31	+5.57	14	+1.03
B	Winter 1969	Common Exam Achievement	41	-2.16	49	-3.74
B	Spring 1968[b]	Grade Achievement	29	+ .200	10	+ .110
B	Spring 1968[c]	Grade Achievement	5	- .080	21	- .276
C	Winter 1969[a]	Common Exam Achievement	20	+3.28	11	-2.55
C	Winter 1969[b]	Common Exam Achievement	16	+4.25	13	+ .57
C	Fall 1967	Grade Achievement	46	- .011	25	+ .128
D	Winter 1969	Common Exam Achievement	29	-1.25	15	-4.46
E	Winter 1969	Common Exam Achievement	77	+2.90	39	+ .67
F	Fall 1967	Grade Achievement	21	+ .052	18	+ .006
G	Fall 1967	Grade Achievement	19	+ .247	9	- .556
H	Fall 1967	Grade Achievement	19	+ .142	15	- .113

[a] Theoretical oriented section.   [b] Control section.   [c] Partial independent study section.

TABLE 2

STUDENT AGE GROUPS VERSUS INSTRUCTOR GROUPS—COMMON EXAMINATION ACHIEVEMENT COMPARISONS, WINTER 1969

Instructor Groups	Age			
	Under 21		21 and Older	
	N	Mean	N	Mean
A, B, and C	171	-1.0	66	+2.0
D and E	108	+1.0	68	- .2

In order to further explore the differences between these instructor groups, the individual specimens on the performance checklist were examined. Items which discriminate are listed in Table 4. Cluster 1 describes instructors A, B, and C as being more likely to apply pressure by forcing students to explain themselves and reach their own conclusions, by lecturing above their heads, and by giving unusually challenging tests which covered much material over a long period of time. This same instructor group was described by cluster 2 as more apt to en-

TABLE 3

STUDENT AGES GROUPS VERSUS INSTRUCTOR GROUPS—OVERALL RATINGS ON PERFORMANCE CHECKLIST, WINTER 1969

Instructor Groups and Criterion Measure	Age			
	Under 21		21 and Older	
	N	Mean	N	Mean
A, B, and C				
Summation of Stimuli Intensity	115	+13.9	44	+34.3
Mean Stimuli Intensity	115	+ .98	44	+2.05
Global Rating	134	2.39	51	2.58
D and E				
Summation of Stimuli Intensity	58	+50.05	38	+58.2
Mean Stimuli Intensity	58	+2.73	38	+2.45
Global Rating	69	+3.12	42	3.18

TABLE 4

FACULTY GROUPS DISCRIMINATING ACHIEVEMENT OF DIFFERENT AGE STUDENTS–COMPARISON OF STIMULI SPECIMENS

| | Faculty Groups | | | | Level of Significance of Difference Between Faculty Groups | |
| | N = 165 A, B, and C | | N = 98 D and E | | | |
	Percent Observed	Mean Intensity	Percent Observed	Mean Intensity	Percent Observed	Mean Intensity
Cluster 1 - More Pressure						
Forced students to qualify, explain, or justify statements and assertations made in class	32	+1.1	10	+ .3	.001	N.S.
Required students to arrive at their own conclusions on class discussion or problems	25	-1.6	2	+2.5	.001	.02
Lectured above students' level of comprehension repeatedly	28	-2.2	5	0	.001	.05
Gave unusually challenging tests which necessitated extensive preparation and thus resulted in definite learning	36	+2.1	16	+2.9	.001	N.S.
Administered tests infrequently, forcing students to cover too much material for a single test	28	-3.1	25	+1.5	N.S.	.001
Cluster 2 - Active Leadership						
Requested and obtained students' questions and reactions	47	+2.2	29	+3.6	.01	.01
Made a dramatic gesture to emphasize a point	22	+2.5	9	+2.7	.01	N.S.
Cluster 3 - Extra Help						
Adjusted his pace to the needs of the class	66	+1.6	85	+4.1	.001	.001
Extended his office hours in order to further assist students	29	+1.4	41	+3.6	.05	.001
Prepared the student for difficulties that might be encountered on an assignment	72	+2.5	79	+3.9	N.S.	.001
Cluster 4 - Laissez-Faire						
Permitted classroom distractions to go unchecked	4	+1.7	21	-1.2	.001	.01
Demonstrated tolerance toward students' ideas even when they conflicted with lectures or course materials	24	+ .7	28	+4.0	N.S.	.001

gage in active leadership by requesting student reaction and by making dramatic gestures. Instructors D and E are depicted in cluster 3 as offering more help by adjusting their pace to class needs, by extending their office hours, and by preparing students for future difficulties. They are further described in cluster 4 as laissez-faire. It appeared they were more apt to let classroom distractions go unchecked

and to tolerate diverse opinions. From these specimen descriptions it might be inferred that older students learn more effectively in an environment which includes pressure and active leadership, but they are not particularly aided by an environment which provides extra help and a _laissez-faire_ atmosphere.

The relationship between rating and achievement provided another interesting avenue for research. In general the correlation of the various performance ratings with common exam achievement was about .30. More important, instructors whose students achieved highest on a common exam were also rated the highest by these students. This is shown in Table 5 and was also demonstrated by Douglass (4).

TABLE 5

COMPARISON OF INSTRUCTORS
BY ACHIEVEMENT AND RATINGS

Instructor	Common Exam Achievement	Summation of Stimuli Intensity	Mean Stimuli Intensity	Global Rating
		Winter 1969		
E	+1.9	+64	+3.1	3.4
C	+1.9	+54	+2.9	3.3
A	+1.5	+41	+2.2	3.0
B	-2.1	-25	-.7	1.4
D	-2.9	+33	+1.6	2.7
Correlation with common exam achievement		.71	.76	.74

Table 6 describes how the highest rated instructor differed from the lowest rated instructor on specific specimen items. These same items also differentiated the top group of faculty nominated for an outstanding teaching award from another random group of faculty at The University of Tennessee (16).

After 2 years of conducting this project several factors stood out with regard to the overall design of the project. There were many confounding variables involved in the effort to describe the learning process. For this reason when a new learning environment is developed or identified, a researcher needs to know if its effects can be generalized over a number of instructors and over time. In order to achieve this generalization the basic unit of analysis must consist of the class section not the individual student. This requires a large number of sections spanning a substantial period of time and including a variety of different instructors. Also required is the use of the same criterion measures over the entire project and this can be difficult when common examinations are used. The final factor that should be considered is the size of the manipulated change

TABLE 6

INSTRUCTOR E VERSUS INSTRUCTOR B–DIFFERENCES ON PERFORMANCE CHECKLIST ITEMS–PERCENTAGE OF STUDENTS UNDER AN INSTRUCTOR OBSERVING EACH SPECIMEN

Item	Instructor E N=65	Instructor B N=61
Pointed out relationships between his fields and other fields of study	75	51
Lectured in a monotone	25	52
Introduced humor to stimulate class interest	97	15
Demonstrated the importance and significance of his subject matter	65	43
Clearly stated the purpose and objectives of the course	80	46
Adjusted his pace to the needs of the class	83	62
Summarized material and showed relationships in a manner which aided retention	71	23
Lectured in a manner which failed to hold class attention	6	48
Refused to explain the basis for his grading system	3	23
Utilized audio or visual aids including blackboard illustrations to clarify lesson materials	82	51
Course assignments remained vague and disorganized	2	26
Lectured in a rambling, disorganized fashion	0	43

in the learning environment. This author feels in order to produce a meaningful impact on learning at the college level, drastic changes that go beyond the normal institutional constraints are required. Rosenbloom (12) concurs that this type of learning environment research be done in research centers which are autonomous from colleges and universities.

FOOTNOTE

1. This is part of a larger study fully reported in a doctoral dissertation entitled "Optimal Learning Environments for Different Types of Students," dated August 1969.

REFERENCES

1. Anderson, H. M., "A Study of Certain Criteria of Teacher Effectiveness," The Journal of Experimental Education, 23:41-71, 1954.

2. Davis, Keith G.; Banning, James H., "Effect of Personality on Programmed Instruction," Research in Education, Index EP-010-541, April 1967.

3. Doty, Barbara A., "Teaching Method Effectiveness in Relation to Certain Student Characteristics," The Journal of Educational Research, 60:363-365, April 1967.

4. Douglass, Linda Grove, "Measures of Teacher Evaluation as Related to Student Achievement," unpublished master's thesis, The University of Tennessee, Knoxville, March 1968.

5. Erickson, Ralph I., "Programmed Learning and Personality Styles at the College Level," The Journal of Educational Research, 60:330-333, April 1967.

6. Goldberg, Lewis R., "Student Personality Characteristics and Optimal College Learning Conditions," Research in Education, Index EP-010-362, February 1967.

7. Jones, R. D., "The Prediction of Teaching Efficiency," The Journal of Experimental Education, 12:85-99, 1946.

8. Katz, Joseph; Sanford, Nevitt, "The Curriculum in the Perspective of the Theory of Personality Development," in Sanford, Nevitt (Ed.), The American College, John Wiley and Sons, Inc., New York, 1962.

9. McKeachie, Wilbert I., "Procedures and Techniques of Teaching: A Survey of Experimental Studies," in Sanford, Nevitt (Ed.), The American College, John Wiley and Sons, Inc., New York, 1962.

10. McKeachie, Wilbert I.; Isaacson, Robert L.; Milholland, John E., Research on the Characteristics of Effective College Teaching, U.S. Department of Health, Education, and Welfare, Office of Education, Cooperative Research Project No. 850, The University of Michigan, Ann Arbor, 1964.

11. Remmers, H. H., "Student Opinion of Teacher Efficiency as a Function of Rank in Class," School and Society, 28:759-760, 1928.

12. Rosenbloom, Paul C. (Ed.), "The Minnesota National Laboratory," in Modern Viewpoints in the Curriculum, McGraw-Hill Book Co., Inc., New York, 1964.

13. Seibert, Warren F., "Linear Program Characteristics," Research in Education, Index EP-010-452, March 1967.

14. Shuford, Emir H.; Massengill, H. Edward, The Relative Effectiveness of Five Instructional Strategies, paper presented for meeting of the National Society for Programmed Instruction, Lexington, Massachusetts, 1967.

15. Siegel, Laurence, "The Contributions and Implications of Recent Research Related to Improving Teaching and Learning," in Milton, Ohmer; Shoben, Edward, Jr., (Eds.), Learning and the Professors, Ohio University Press, Athens, 1968, pp. 136-157.

16. Whitlock, Gerald H., "Criteria for Selecting Recipients of the Outstanding Teaching Award," Department of Industrial Management, The University of Tennessee, Knoxville, 1966, unpublished.

17. Whitlock, Gerald H., "Criteria of Teaching Effectiveness and Problems of Measurement," in Milton, Ohmer (Ed.), A Conference on Improving Arrangements for Learning in Tennessee Colleges and Universities, The University of Tennessee, Knoxville, December 1964, pp. 15-18.

ON IMPROVING THE PERFORMANCE

OF CLASSIFICATION TECHNIQUES[1]

P. JOSEPH PHILLIP
University of Detroit

ABSTRACT

This study is based on the antecedent characteristics of 790 students enrolled in a professional degree program. It investigates two approaches toward improving the performance of classification techniques. The first approach focuses attention on the selection of predictors using scaled distances and intra-population correlations instead of t-ratios. The second involves the development of a Bayesian Taxonomic Procedure in addition to the traditional technique, the Linear Discriminant Function. It was found that reliance on t-ratios alone will not guarantee the selection of the best predictor set. As for the relative performance of the classification techniques, no significant difference was observed. However, certain hypotheses concerning the strength and weakness of each technique emerged from the study.

THE LAST four decades have witnessed a proliferation of studies designed to predict academic success defined in terms of course grades. The problem was attacked from every conceivable angle; yet predictive validity soon reached an unsatisfactory plateau (10). A possible explanation for this lack of progress may be sought in the nature of the criterion which is almost always defined in terms of grades. A criterion so defined is subject to considerable instability, a part of which stems from the vagaries of grading practices (6). Exclusive emphasis on grades is also open to question from the standpoint of construct validity. Current research indicates that the intellectual factor is but one factor of academic success, and that there are other factors orthogonal to intelligence (4). Finally, if we view the educational process within a broader context, as we must, the question arises as to how far grades correlate with later occupational success. Investigations in this area indicate no significant correlation (13). It is felt that the broader criterion of success versus failure is less open to the criticisms raised above. Of course, this will not eliminate the vagaries of grading practices. But its impact on the criterion is likely to be less decisive.

For the prediction of group membership the Linear Discriminant Function (LDF) is generally regarded as the most appropriate statistical technique (14). There have been a number of applications of this technique to taxonomic problems in education.

Although these studies represent an advance in the sense that they have caused some shift in emphasis away from grade point average (GPA) as the be-all and end-all of the educational process, the results have not always been entirely satisfactory. Even the reported performance of the LDF may be an over-estimation of what it can really accomplish (5). It will be instructive to explore ways in which better prediction of group membership could be achieved. The present study investigates two approaches to this problem.

The first approach focuses attention on the selection of predictor variables for the LDF. Unlike regression analysis, no systematic procedures for the selection of variables have been developed for the LDF. True, some guidelines are available in the works of Rao (12), Lubischew (9), and others. But the contributions of these writers have not been woven into an integrated pattern within the framework of a general theory of LDF. Consequently, a researcher, confronted with numerous variables to choose from, has to "play it by ear." The usual practice is to select variables by a t-test at some level of significance. That this may not be the most efficient procedure has been demonstrated recently by Cochran (2).

The second approach involves the application of a classification technique other than the LDF. Many variables in educational and psychological research are qualitative. Despite the fact that these situations

represent departures from the classical theory of LDF, one often finds that a linear function of a set of qualitative variables is used for classification purposes. Even when the variables are continuous, the assumption that the k populations have a multivariate normal distribution with equal dispersion on p measurements is rarely met. A classification technique which does not involve the rigorous assumptions of the LDF has, therefore, much to commend it.

PROCEDURE

Sample, Definition of Terms, and Variables

The sample for this study (11) consisted of students enrolled in the Doctor of Dental Surgery program at the University of Detroit School of Dentistry from 1953-1963. The sample size was 790, of which a sub-sample of one hundred students, drawn at random, was set aside for cross-validation. The criterion is defined as follows: Success Group = Those who graduated on schedule; Failure Group = Those who withdrew voluntarily or involuntarily, or failed to graduate on schedule.

In all, thirty-five predictor variables, including biographic, scholastic, and aptitude (Dental Aptitude Test) variables were analyzed.

Selection of Predictors

Cochran (2) has shown that if the population could be assumed to be multivariate normal, the taxonomic power of a variable, say X_i, is given by the scaled distance, d_i, where

$$d_i = \frac{(\overline{X}_{is} - \overline{X}_{if})}{\sigma_{ip}} \quad (1)$$

where $\overline{X}_{is}$ and $\overline{X}_{if}$ are the means of variable X_i in populations S and F respectively; and σ_{ip} is the pooled within-group standard deviation. If the samples are not too small, d_i would provide an estimate of misclassification when variable X_i alone is used for classification. This is equal to the probability that a random normal deviate will exceed $d_i/2$, and so can be obtained from a table of normal distribution. For example, if $d_i = 2.56$, $z = 2.56/2 = 1.28$, and the probability of misclassification is 10 percent; when a second variable is added to a variable already selected, the combined discriminatory power of the variable depends upon (i) the discriminatory power, d, of the individual variables, and (ii) the intra-population correlation, ρ_{ij}.[2] More specifically, the combined squared distance, d^2_c, of variables X_i and X_j ($d_i > d_j$) will be determined as follows:

CASE   FORMULA

A. d_j Significant
   ρ_{ij} Significant
$$d^2_c = d^2_i + \left[(d_j - \rho_{ij}d_i)^2/(1-\rho^2_{ij})\right] \quad (2)$$

B. d_j Significant
   ρ_{ij} Not Significant
$$d^2_c = d^2_i + d^2_j \quad (3)$$

C. d_j Not Significant
   ρ_{ij} Significant
$$d^2_c = d^2_i + d^2_i \left[\rho^2_{ij}/(1-\rho^2_{ij})\right] \quad (4)$$

The implications of these formulas are worth noting: A negative ρ_{ij} always increases the combined discrimination, while a positive ρ_{ij} is, in most cases, harmful (Case A). If $\rho_{ij} = 0$, the second variable contributes to the extent of its discriminatory power (Case B). A covariate or a "suppressor variable" would improve discrimination if (a) its correlation with the variable already selected (positive or negative) is high and (b) the latter itself has sizable discriminatory power (Case C).

When more than two variables are involved, the formula for computing the d^2_c gets quite complex. It is clear, however, that if a variable does not have a significant d_i value or a significant ρ_{ij} with at least one of the discriminants, it could not, in any way, contribute toward discrimination. Such variables were eliminated, and the deletion procedure suggested by Rao (12:249-255) was applied to the remaining variables.

The Linear Discriminant Function

When the dependent variable is a dichotomy, like success versus failure, what is needed is a technique which predicts group membership by focusing attention on the differences between groups rather than the differences within a single group. Fisher's LDF does just that. If we have p measurements, X_1, X_2 . . . . X_p on two groups of sample sizes, N_1 and N_2, the LDF provides maximum separation of these two groups by maximizing the ratio of the difference between the specific means, $\overline{X}_1$ and $\overline{X}_2$ to variations within groups, or

$$\text{Max } G = \frac{(\overline{X}_1 - \overline{X}_2)^2}{\frac{2}{\sum_{i=1}} \frac{ni}{\sum_{j=1}} (X_{ij} - \overline{X}_i)^2} \quad (5)$$

Fisher does not give the rationale for restricting his solution of the problem to linear equations, but subsequently Welch (15) and Hodges (8) showed that when the observations arise from normal populations with the same variance-covariance matrix, the function has optimal properties.

It is conventional to test the effectiveness of the LDF by the generalized distance function, D^2, using the F ratio:

$$F = \frac{N_1 N_2 (N_1 + N_2 - p - 1) \cdot D^2}{p (N_1 + N_2) (N_1 + N_2 - 2)} \quad (6)$$

with p and $N_1 + N_2 - p - 1$ degrees of freedom. If the F test is significant at a suitable level, a discriminant equation of the following form is constructed:

$$y = b_1 X_1 + b_2 X_2 + \ldots + b_p X_p \quad (7)$$

An individual's discriminant score, y, is computed by plugging the measurements made on him into equation (7) and he is assigned to Group 1 or Group 2 according to whether his discriminant score falls above or below a critical value established as follows:

$$M^*_y = \frac{\left[\overline{y}^2_2 - \overline{y}^2_1 + (2\sigma^2 y \log_e P_1/P_2)\right]}{2(\overline{y}_2 - \overline{y}_1)} \quad (8)$$

Figure 1

THE TREE DIAGRAM
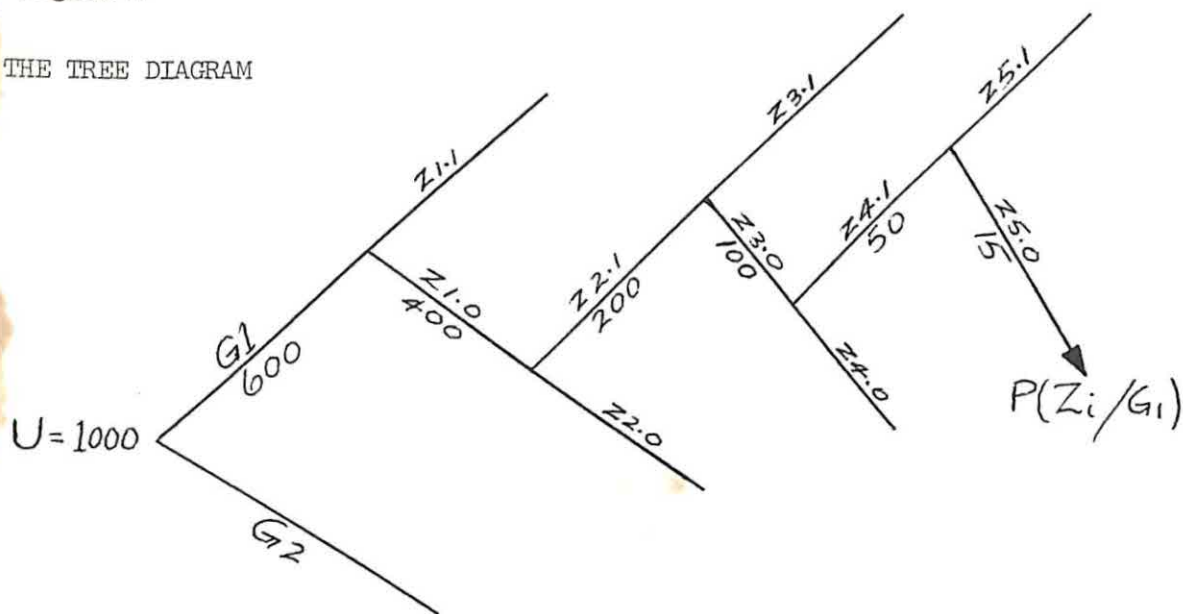
$$P(G_1) \quad = \frac{G_1}{U} \quad = \quad \frac{600}{1000} \quad = \quad \underline{0.600}$$

$$P(Z_i/G_1) \quad = P(z_{1.0} \frown z_{2.1} \frown z_{3.0} \frown z_{4.1} \frown z_{5.0}) \quad = \frac{Z_i}{G_1}$$

$$= \frac{400}{600} \times \frac{200}{400} \times \frac{100}{200} \times \frac{50}{100} \times \frac{15}{50} \quad = \frac{15}{600} = \underline{0.025}$$

$$P(G_1).P(Z_i/G_1) = P(G_1 \frown z_{1.0} \frown z_{2.1} \frown z_{3.0} \frown z_{4.1} \frown z_{5.0}) = \frac{Z_i}{U}$$

$$= \frac{600}{1000} \times \frac{400}{600} \times \frac{200}{400} \times \frac{100}{200} \times \frac{50}{100} \times \frac{15}{50} \quad = \frac{15}{1000} = \underline{0.015}$$

where M^*_y is the critical value, $\bar{y}_1$ is the mean of the larger group, $\bar{y}_2$ is the mean of the smaller group, σ^2_y is the overall variance of y, P_1 is $N_1/(N_1+N_2)$ and P_2 is $N_2/(N_1+N_2)$. When $P_1=P_2$, the terms within simple brackets vanish so that equation (8) reduces to:

$$M^*_y = \frac{(\bar{y}_2 + \bar{y}_1)}{2} \qquad (9)$$

which will be readily recognized as the middle value of the means (in discriminant scores) of Group 1 and Group 2.

The Bayesian Taxonomic Procedure

Our point of departure is the concept of prior probabilities for the Success Group (G_1) and the Failure Group (G_2). This is simply the "base rates" or the average proportion of successes and failures observed over a period of time in the student population. Analysis of the attribute pattern (Z_i) of students within G_1 and G_2 yields conditional probabilities, that is, the probabilities of possessing attribute pattern (Z_i), given that an individual belongs to G_i. These conditional probabilities can be used in conjunction with the prior probabilities to derive the posterior probabilities of belonging to G_1 and G_2 by application of the Bayes Theorem:

$$P(G_1/Z_i) = \frac{P(Z_i/G_1) \cdot P(G_1)}{P(Z_i/G_1) \cdot P(G_1) + P(Z_i/G_2) \cdot P(G_2)} \qquad (10)$$

$$P(G_2/Z_i) = \frac{P(Z_i/G_2) \cdot P(G_2)}{P(Z_i/G_2) \cdot P(G_2) + P(Z_i/G_1) \cdot P(G_1)} \qquad (11)$$

where $P(G_i/Z_i)$ is the posterior probability of belonging to G_i, given that an individual has the attribute pattern vector Z_i. Here $Z_i = z_{1 \cdot j}, z_{2 \cdot j}, z_{3 \cdot j}, z_{4 \cdot j}, z_{5 \cdot j}$ where the first subscript identifies the attribute, and j=1 denotes the presence and j = 0 denotes the absence of the attribute. [3] $P(Z_i/G_i)$ is the conditional probability and $P(G_i)$ is the prior probability.

The derivation of all measures required for formulas (10) and (11) is illustrated in the tree diagram (Figure 1) where hypothetical figures are used.

FINDINGS

Selection of Predictors

Of the thirty-five variables included in the analysis, nine were eliminated because they lacked discriminatory power as well as intra-population correlation with any of the discriminants. The predictors finally selected for the classification models are listed in Table 1.

It was found that the first two variables (X_1 and X_{24}) yielded the largest combined distance than all other pairs of variables; and the first three variables (X_1, X_{24}, and X_{34}) yielded a combined distance

larger than all other combinations of three variables. In fact, the combined distance of these three variables was found to be 1.081, while the combined distance of the triad with the largest t-ratios (X_6, X_1, and X_{34}) was only 0.467. This would seem to suggest that exclusive reliance on t-ratios does not always guarantee that the most productive variables will be selected because intra-population correlation is ignored. Another drawback of a selection procedure based on t-ratios alone is the danger of discarding covariates. In the present study, although none of the covariates qualified for selection, at least two came close to being selected.

The Effectiveness of Models

A model which classifies subjects as members of the Success Group and the Failure Group produces two types of errors or misclassifications. Some of those classified as successes may in fact be failures, the "false positives." Some of those classified as failures may in fact be successes, the "false negatives." A classification strategy may be so devised as to minimize "false positives," "false negatives," or total misclassifications. The research problem usually dictates which of these should be minimized. The present research is concerned with the selection of a fixed number of candidates from a large applicant pool. Since selection may be regarded as a special case of classification where the cost of "false negatives" is zero or negligible, it becomes apparent that minimization of "false positives" has to be the primary aim.

TABLE 1

THE PREDICTORS SELECTED
FOR CLASSIFICATION MODELS

Predictor	Scaled Distance	t-Ratio
X_1: Location of Residence	0.329	3.500
X_{24}: Manual Average	0.327	3.473
X_{34}: Spacial Relations	0.329	3.500
X_9: Chemistry Honor Point Average	0.326	3.4?
X_6: Number of Colleges Attended	0.342	3.6?

Having formulated the classification strategy, the performance of the model may be evaluated against the a priori strategy, that is, the prevailing strategy not using the model. For the present study, the a priori strategy is the current selection procedure, and the best estimate of the efficiency of this strategy is provided by the "base rates," that is, the average proportion of successes and failures that occurred when this strategy was in force. For the University of Detroit Dental School, the base rates for the success and failure groups are 0.7924 and 0.2058 respectively. A classification model is adjudged superior to

the a priori strategy if it significantly changes the base rates in favor of the success group.

For the present research, two classification strategies were formulated. Strategy 1 assigned individuals of the cross-validation group in accordance with a decision rule which imputes equal cost to both types of misclassifications. [4] This was accomplished by applying formula 8 to determine the critical value for the LDF, and by setting $P(G_1) = 0.7924$ and $P(G_2) = 0.2058$ for the Bayesian Taxonomic Procedure (BTP). The resulting decision rules are listed in Table 2.

TABLE 2

DECISION RULE FOR STRATEGY 1

The Linear Discriminant Function

1. If $y \geq 2.538$ Assign to G_1

2. If $y < 2.538$ Assign to G_2

The Bayesian Taxonomic Procedure

1. If $P(G_1/Z_i) \geq 0.500$ Assign to G_1

2. If $P(G_1/Z_i) < 0.500$ Assign to G_2

When assignments were made in accordance with these decision rules both models assigned practically all members of the cross-validation group to the success group (G_1). This result, disappointing though it may seem, underlines the problem intorduced by lopsided prior probabilities. Consider, for example, a binomial population with $P_1 = 0.99$ and $P_2 = 0.01$. Since a strategy which assigns everyone to P_1 will result in a misclassification rate of only 1 percent, any alternative strategy should possess uncanny accuracy to be justified in terms of incremental validity.

Strategy 2 assigned individuals in accordance with a decision rule which assigned higher cost to "false positives." This was accomplished by applying formula 9 to determine the critical value for the LDF, and by setting $P(G_1') = P(G_2') = 0.500$ for the BTP. The resulting decision rules are listed in Table 3.

When assignments were made in accordance with these decision rules, the proportion correctly assigned to the success group by the LDF and the BTP were 0.8933 and 0.9077 respectively. While these proportions do not differ from one another statistically, they represent a significant improvement over the a priori strategy which produces a success rate of 0.7942.

Significance for Future Research

The present study reveals no significant difference between the performance of the LDF and the BTP. This is true regardless of the strategy under which assignments are made, or the criterion against which performance is evaluated. Perhaps the nonsignificant difference is due to the hybrid nature of the predictors used to develop the models. From the stand-

TABLE 3

DECISION RULE FOR STRATEGY 2

The Linear Discriminant Function

1. If $y \geq 3.992$ Assign to G_1

2. If $y < 3.992$ Assign to G_2

The Bayesian Taxonomic Procedure

1. If $P(G_1'/Z_i) \geq 0.500$ Assign to G_1

2. If $P(G_1'/Z_i) < 0.500$ Assign to G_2

point of the practitioner, the LDF and the BTP have their strong and weak points. For the LDF to have optimal properties, the following conditions must be met: the predictor measures should be based on random samples drawn from multivariate normal populations with homogeneous dispersions, and the relationship between the predictors and the criterion should not depart significantly from linearity. These conditions are rarely met in practice. The BTP involves the reduction of predictors into a few states, preferably dichotomies, a procedure which may be advantageous when the relationship is sharply nonlinear, but which entails considerable loss of information content and degrees of freedom. It would therefore seem intuitively that the effectiveness of the models depends a great deal upon the quality of data. It would be insightful to test this hypothesis by developing both models from one set of data which does not violate these requirements and another set which does violate these requirements. The analysis performed on the first set would show to what extent the loss of information content and degrees of freedom impairs the performance of the BTP; and the analysis performed on the second set would show to what extent failure to meet the underlying assumptions of LDF impairs the performance of that model.

Certain features peculiar to the BTP began to show up during the course of the research. They are summarized below:

a. Since the Bayesian scheme with k dichotomized variables will generate 2^k mutually exclusive and collectively exhaustive attribute vectors, a large volume of data is a sine qua non.

b. It is essential that the number of predictors be limited to, say, four or five, for an increase in predictors is accompanied by a tremendous proliferation of attribute vectors and rapid decrease in the number of observations per vector. Furthermore, when the predictors are increased, the attribute pattern matrix becomes quite unweildy. For example, ten predictors generate a 1024x10 attribute pattern matrix.

FOOTNOTES

1. The author wishes to acknowledge the help of his major advisor, Dr. William Reitz, in carrying out this investigation.

2. This measure is the correlation coefficient computed with pooled variance and covariance, and the sign determined as follows: Let $r_{\bar{x}_i\bar{x}_j}$ be the correlation between population means. If $\bar{X}_i$ is larger or smaller than $\bar{X}_j$ in both populations, we say that $r_{\bar{x}_i\bar{x}_j}$ is positive. If $\bar{X}_i$ is larger than $\bar{X}_j$ in one population and the reverse is true in the other population, we say that $r_{\bar{x}_i\bar{x}_j}$ is negative. If $r_{\bar{x}_i\bar{x}_j}$ and the computed correlation coefficient have like signs, ρ_{ij} will be positive, if they have unlike signs, ρ_{ij} will be negative. See Lubischew (9).

3. In order to develop the attribute vectors, the variables used for the Bayesian model were dichotomized in a manner that minimized total misclassifications (7).

4. This strategy may also be defined as one that minimizes total misclassification. See Cooley; Lohnes (3:137) and Birnbaum; Maxwell (1:252).

REFERENCES

1. Birnbaum, Allen; Maxwell, Albert E., "Classification Procedure Based on Bayes's Formula," in Cronbach, L. E.; Glesser, G. C. (Eds.), Psychological Tests and Personnel Decisions, University of Illinois Press, Urbana, 1965.

2. Cochran, William G., "On the Performance of the Linear Discriminant Function," Technometrics, 6:179-190, 1964.

3. Cooley, William W.; Lohnes, Paul R., Multivariate Procedures for Behavioral Sciences, John Wiley, New York, 1962.

4. Finger, John A., Jr., "Academic Motivation and Youth-Culture Involvement: Their Relationships to School Performance and Career Success," School Review, 74:177-195, 1966.

5. Frank, Ronald E. and others, "Bias in Multiple Discriminant Analysis," Journal of Marketing Research, 2:250-258, 1965.

6. Frederiksen, Norman; Schrader, William B., Adjustment in College. Educational Testing Service, Princeton, New Jersey, 1951.

7. Guttman, Isaiah; Raju, Namburi S., "A Minimum Loss Function as Determiner of Optimal Cutting Score," Personnel Psychology, 18:179-182, 1965.

8. Hodges, Joseph L., Jr., Discriminatory Analysis. Randolph AFB, Texas, 1955.

9. Lubischew, Alexander A., "On the Use of a Discriminant Function in Taxonomy," Biometrics, 18:455-477, 1962.

10. Michael, William B., "Measurement and Prediction in College Admission Process: Some Possible Directions for Future Research," Educational and Psychological Measurement, 25:55-72, 1965.

11. Phillip, P. Joseph, "A Comparison of the Relative Effectiveness of the Linear Discriminant Function and the Bayesian Taxonomic Procedure: A Case Study of a Dental School," unpublished doctoral dissertation, Wayne State University, College of Education, Detroit, Michigan, 1969.

12. Rao, C. Radhakrishna, Advanced Statistical Methods in Biometric Research, John Wiley, New York, 1952.

13. Simonds, Rollin H., "College Majors, Grades Versus Business Success," Business Topics, 14:7-12, 1966.

14. Tatsuoka, Maurice; Tiedeman, David V., "Discriminant Analysis," Review of Educational Research, 24:402-420, 1954.

15. Welch, Bernard L., "Notes on Discriminant Functions," Biometrika, 31:218-220, 1939.

HIGHER EDUCATION ADMINISTRATION
STUDENTS' PERCEPTION OF ESTABLISHING
A COMMUNITY COLLEGE

JAMES T. RANSON
West Virginia University – Kanawha Valley Graduate Center

ABSTRACT

The recency of the community college movement has prevented large scale organization and dissemination of knowledge concerning the processes and steps required to establish a community college. There is, therefore, a need for means to incorporate contemporary evidence into the curriculum of prospective administrators of higher education. The purpose of this study was to compare Higher Education Administration students' pecep-tions of a logical sequence of events for establishing a community college with a sequence of events for establishing a community college as determined by practice. The Edwards' matched-pair technique was used to rank the perceptions of the students. A sequence of twenty-nine events was selected from the literature. Each event was matched with every other, thus providing 406 decision situations. From these decisions an ordering was obtained. Using the Spearman rho correlation technique, a coefficient of 0.82 was found between the sequence of events taken from the literature and rank ordering of these events as determined by the Edwards' matched-pair technique.

HOW TO provide education of a useful kind for thousands of youngsters from Florida to Washington and from Arizona to New England is a common question being asked. The answer to this question being proposed in many areas is the community college. For example, a bond issue providing $47,200,000 for developing community colleges was passed by a 4 to 1 margin by voters of St. Louis, Missouri, and across the state Kansas City passed a $25 million bond issue for the same purpose. The state legislature of Connecticut is backing plans for a statewide network of community colleges (3).

In the fall of 1965 alone, eight junior colleges opened in Alabama, Arizona, California, Connecticut, Florida, Michigan, Minnesota, Nebraska, North Carolina, Oregon, Pennsylvania, Texas, Virginia, and Washington. Alabama started 11 and North Carolina 10 at the same time (3:1).

One can hardly deny that the community college is having a tremendous, if not dramatic, impact on higher education in the United States. A truly American innovation, the community college impact, is presenting and shall continue presenting a tremendous challenge for higher education administrators. And more particularly this challenge will no doubt confront those students of higher administration who will be assuming leadership positions in these new institutions within the near future.

The recency of the community college movement has inhibited large scale organization of knowledge concerning sequences of events for establishing community colleges. For example, the work of Stivers (5) and Benson (1) was not available until 1962. There is therefore, a need for more knowledge concerning the application of techniques for systematizing evidence for use in curricula for training community college administrators. One way to help meet this need might be the matched-pair technique developed by Edwards (2) to create decision situations from evidence by practitioners in the community college community. There is some information concerning the establishment of new community colleges. For example, Benson (1) has developed a Program Evaluation Review Technique (PERT) network for such a task and Stivers (5) has recorded a sequence of events functional for establishing a new community college. If information of this type could

be conveniently incorporated in training curricula for prospective higher education administrators, it should enhance the chances for an orderly process for community college development.

The purpose of this study was to compare Higher Education Administration students' perceptions of a logical sequence of events for establishing a community college with a sequence of events for establishing a community college as determined by acutal practice.

METHODOLOGY AND PROCEDURES

Edwards describes the logic of this study (2).

Let us suppose that we have ten objects of the same size but of differing weights and that we wish to arrange the objects from the lightest to the heaviest. We could easily place each object on a scale, read the pointer on a dial, and record the measure of weight. On the basis of our observations, the objects could then be arranged in order from the lightest to the heaviest. Let's suppose that a scale for weighing the objects is not available. Instead of weighing the objects, we would present them to individuals and ask them to make judgments about the respective weights of the objects.

The scale that we use in weighing objects we call a physical scale, and the ordering of the object in terms of measured weights is said to be on a physical continuum. The ordering of the objects upon the basis of judgments is said to be on a psychological continuum (2:19).

The assumption was made that the sequence of events identified by Stivers (5) defined a physical continuum, and that the sequence of events as judged by the group in this study defined a psychological continuum.

The sequence of events as documented by Stivers (5) was arranged according to the event requiring the most lead time to the events requiring the least lead time. For example, the event requiring the most lead time was establishment of the need for a new community college. For purposes of measurement this event was assigned number one. All events were ordered from one through twenty-nine. This operationally defined the physical continuum.

The ordering of events via the matched-pair technique operationally defined the psychological continuum. Each of the events in perceptual sequence was assigned the same number as those in the physical sequence. However, they were not ordered as such. The ordering of the events for the perceptual sequence was secured by responses from the matched-pair comparison instrument.

An incidental sample of twenty-eight students of Higher Education Administration at the Ohio State University assumed to be a random sample from a population of individuals who will become administrators of community colleges, were asked to respond to the instrument during the winter quarter of 1966. One hour of time was allotted and all Ss responded to all items. Sample responses were then submitted to the Ohio State University computer center for processing. (A computer program for this research was designed during the summer quarter of 1965 at the Ohio State University.)

PRESENTATION OF DATA

Table 1 presents the results of this study. It depicts the comparison of the physical continuum of events with the psychological continuum as defined by Ss' perceptions of events that take place in establishing a new community college. Since the level of measurement was ordinal a Spearman rho correlation test was applied to the two event sequences (4). The correlation between the physical and perceptual sequences was 0.82.

Using the number of months from the starting date that each event took place along the physical continuum as units of measure, interval measurement can be assumed. On this assumption a Pearson correlation of 0.834 with a standard error of 0.056 was obtained.

Statistical interpretation can be made on the assumption that the group was a random sample; the null hypothesis would state that the two event sequences were uncorrelated. A student t calculated from the Pearson correlation coefficient of 0.834 yielded a value of 7.846. For significance at the .01 level 2.77 was required. Therefore, the two variables were statistically significantly correlated.

DISCUSSION

The purpose of this study was to compare Higher Education Administration students' perceptions of a logical sequence of events for establishing a community college with a sequence of events for establishing a community college as determined by actual practice.

If one goal of instruction is to provide information to a student in a setting which generates data for instruction, the matched-pair technique is quite useful. In this study a sequence of events recorded as functional in establishing a community college provided a base for comparison and provided 406 decision situations for a group of prospective community college administrators. The data generated defined some interesting instructional areas. For example, evidence from the field indicates that establishing the need for a community college should occur first while the collective judgment of the group indicated that cost estimates should occur first. Field evidence indicates that estimating enrollment and growth is thirteenth in the sequence of events, while the students perceived this as fifth. Evidence from the field indicates that ordering instructional equipment and supplies is eighteenth in the sequence of events, while the students perceived this as seventh. Field evidence indicates that setting up faculty offices is twenty-third in the sequence, while the students perceived this as eleventh.

The differences discussed above are attributable to the less than perfect correlation of 0.86. Ostensibly, the high correlation would lead one to conclude that there is high agreement and, therefore, a great deal of confidence could be placed in the student perceptions. However, the discrepancies noted above

TABLE 1

SEQUENCE OF EVENTS AS THEY OCCURRED IN ESTABLISHING A NEW COMMUNITY COLLEGE AND AS THEY WERE PERCEIVED AS OCCURRING BY STUDENTS OF HIGHER EDUCATION ADMINISTRATION

Time Before Opening Day	Physical Sequence Ranking	Students' Perceptual Sequence Ranking	Ranking Difference
24 months	1. Establish the need	4. Estimate costs	-3
22 months	2. Obtain legal authorization	1. Establish the need	1
20 months	3. Determine financial support	3. Determine financial support	0
20 months	4. Estimate costs	2. Obtain legal authorization	2
18 months	5. Appoint a president	13. Estimate enrollment and growth	-8
18 months	6. Begin intensive publicity	8. Determine location	-2
18 months	7. Engage an architect	18. Order instructional equipment/ supplies	-11
15 months	8. Determine location	5. Appoint a president	3
12 months	9. Select key administrators	14. Set up a budget	-5
12 months	10. Provide secretarial help	10. Provide secretarial help	0
12 months	11. Purchase office equipment/ supplies	23. Set up faculty offices	-12
12 months	12. Solicit scholarships and books	7. Engage an architect	5
9 months	13. Estimate enrollment and growth	17. Select instructional staff	-4
9 months	14. Set up a budget	6. Begin intensive publicity	8
6 months	15. Determine curriculum, issue catalog	9. Select key administrators	6
6 months	16. Engage remaining administrators	19. Order textbooks/supplies	-3
6 months	17. Select instructional staff	11. Purchase office equipment/ supplies	6
5 months	18. Order instructional equipment/ supplies	16. Engage remaining administrators	2
2 months	19. Order textbooks/supplies	15. Determine curriculum/issue catalog	4
2 months	20. Employ non-instructional help	24. Set up classrooms	-4
1 month	21. Set up bookstore	12. Solicit scholarships and books	9
1 month	22. Set up lunchroom	21. Set up bookstore	1
1 month	23. Set up faculty offices	26. Process faculty/student handbooks	-3
½ month	24. Set up classrooms	27. First faculty meeting	-3
½ month	25. Arrange for vending machines	20. Employ non-instructional help	5
½ month	26. Process faculty and student handbooks	22. Set up lunchroom	4
½ month	27. First faculty meeting	25. Arrange for vending machines	2
¼ month	28. Registration	28. Registration	0
0	29. Opening Day	29. Opening Day	0

could well lead to difficulty if the physical sequence
was accepted as the only valid sequence.

The discrepancies could well be indications of issues which could be raised concerning the validity of the physical continuum. Because there is no evidence to support the perceptual sequence this should not lead to the conclusion that it is invalid. Research and/or practice could well substantiate the psychological continuum as a valid description of a sequence of events for establishing a community college.

Certainly the collective judgment of this group of prospective community college administrators concerning a sequence of events determined to be functional in setting up a community college was remarkably valid, and the Edwards' matched-pair comparison is a feasible method for incorporating field evidence into instruction and evaluation of instruction.

REFERENCES

1. Benson, Ellis M., "A Time and Sequence Analysis of Critical Steps in the Establishment of California Public Junior Colleges," unpublished doctor's thesis, University of California, Los Angeles, 1962.

2. Edwards, Allen, Techniques of Attitude Scale Construction, Appleton-Century-Crofts, Inc., New York, 1957.

3. Footlick, Jerrold K., The National Observer, November 29, 1965.

4. Kerlinger, Fred N., Foundation of Behavioral Research, Holt, Rinehart, and Winston, Inc., New York, 1964.

5. Stivers, Earl R., "A Junior College Timetable," Overview, 3:38-40, October 1962.

A REINFORCEMENT ANALYSIS OF THREE-MAN TEAM PERFORMANCE IN A PSYCHOLOGY COURSE

JON E. ROECKELEIN [1]
Mesa Community College

ABSTRACT

Student performance was measured in an introductory psychology course organized into four 3-man "series" and four 3-man "parallel" teams. Six hypotheses were set up to be tested in a situation where the primary dependent variables were responses made on quizzes by students performing alone, and students performing as part of a team. The hypotheses dealing with intra- and inter-team functioning were not confirmed. Hypotheses concerned with overall individual proficiency levels were supported by substantial gains made on post-team versus pre-team tests. It was concluded that reinforcement theory concepts were extrapolated successfully from the experimental laboratory to the college classroom when one considered the specific educational and behavioral objectives which the course of instruction sought to accomplish.

SOME INVESTIGATORS (e.g., 2), study group behavior in the same manner in which individual behavior is usually studied with operant conditioning procedures. The primary unit in these investigations is the behavior of the team rather than the individual and emphasizes the feedback and reinforcement contingencies that are produced as a function of the "group environment." Using this approach, Glaser and Klaus (2) formed "series" and "parallel" teams to study how group feedback, which comprised the reinforcing event, could be contingent upon combined individual performances. Briefly, in a series team, if one member responds incorrectly, no reinforcing feedback is presented to other members even though they may have made correct responses. In a parallel team, on the other hand, a correct response by one or more members can produce a correct team response.

The purpose of the present study was to attempt an extrapolation from studies conducted in the laboratory (e.g., 2) to studies conducted in a formal educational setting by applying the methodology involving 3-man series and 3-man parallel teams to the relatively complex environment of the college classroom. Accordingly, within the general framework of a reinforcement analysis, the following six hypotheses were set up to be tested in a classroom learning situation:

Hypothesis 1. The structure of a series team, and the nature of information processing within a series team, will result in an overall team performance increment; in addition, individual member performance levels will be higher for individuals when they are members of a parallel team.

This characteristic of a series, or non-redundant membership, team may be explained by the reinforcing condition which permits reinforcement only for correct member responses. In this situation, error responses should rapidly extinguish.

Hypothesis 2. The structure of a series team will result in the application of more "social pressure" to certain individual member(s) by other members on the team as compared with the "social pressure" found in a parallel team.

This hypothesis is based upon the relative importance which the present method attaches to the consistent high-level performance required from each member of a series team (i.e., non-redundant membership) and was tested by comparing the number of individuals absent from class for series versus parallel teams.

Hypothesis 3. The structure of a parallel team, and the nature of information processing within a parallel team, will result in an overall team performance decrement.

This characteristic of a parallel, or redundant membership, team may be explained by the reinforcing condition which permits aperiodic reinforcement for incorrect member responses. It is assumed in this situation that error responses will not readily extinguish.

Hypothesis 4. For a parallel team consisting of redundant members who have initially divergent proficiency levels, one very high and the other low, team performance will become primarily a function of the more proficient member and the contributions of the poorer member will become increasingly small.

This hypothesis was tested by analyzing the individual member's performance on sequences of exams or quizzes.

Hypothesis 5. All members of all teams in all sections of the class will perform individually at higher proficiency levels on psychology exams at the end of the course than at the beginning of the course as a result of holding membership in both series and parallel teams.

This hypothesis was tested by comparing intra-member proficiency scores made after the first week of the course with those made after the last week of the course.

Hypothesis 6. All members of all teams in all sections of the class will develop individually more critical ways of solving problems and will develop individually more rigorous study habits and attitudes.

This hypothesis was tested by comparing intra-member scores on two separate forms of the Watson-Glaser Critical Thinking Appraisal (CTA) and the Brown-Holtzman Survey of Study Habits and Attitudes (SSHA).

METHOD

Subjects and Procedure[2]

Twenty-four students in an introductory psychology course (summer session) were divided into 3-man teams comprising two sections of teams: four series and four parallel teams. Teams and sections were formed 1 week after the beginning of the 5-week course which met for 1 hour, 45 minutes each day for 5 days a week. The first week was devoted to testing individual students to obtain individual proficiency scores which were used when forming groups to insure initial homogeneity across teams and sections. During the first week, students were informed that the two primary objectives of the course were: (1) To provide the student with selected facts, principles, and concepts of general psychology; and (2) To increase the students' proficiency in taking multiple-choice and fill-in type examinations. Two forms each of the CTA and SSHA were administered to the students, one during the first week and one during the last week of the course. The same final exam, covering the entire course work in psychology, was also given during the first and last weeks. The regular testing materials consisted of multiple-choice and fill-in type questions drawn from the textbook material assigned. All questions were chosen from an exam pool used in other introductory psychology courses at that school. Each quiz used in the course consisted of ten questions. Response bias across quizzes was carefully controlled; that is, all quizzes had an overall equivalent level of difficulty based upon previous usage of questions. During the course, no formal lectures were scheduled by the instructor; however, a table labeled "Lecture Table" was available to all teams throughout the course and a team could meet with the instructor for 5-6 minutes at a time for short lectures on any material which was unclear to the team concerning a particular assignment.

The overall design of the present study may be classified as "quasi-experimental, time-series" (1). The intention was to provide situations where periodic measurement instruments (quizzes) could be administered to the group and certain experimental changes (series and parallel team conditions) could be applied at various places throughout the experimental space (college course in introductory psychology).

RESULTS AND DISCUSSION

Table 1 contains the results of three analyses involving the total number of error responses produced by series and parallel teams (all students served under both series and parallel team conditions for an equivalent amount of time). Two of the analyses were made by splitting each type of team, series or parallel, in half and comparing the halves (Series A, Series B, or Parallel A, Parallel B) through the application of statistical tests. Since there were an odd number of days in the total number of days used in the split-half method, it was decided arbitrarily that team performances from the middle day separating the two halves would be excluded from analysis; thus, the middle day was considered to be a "buffer" day which conveniently divided intra-team performances for the series and parallel conditions. The third analysis was made by comparing the total series team responses of a section with total parallel team responses of the same section, buffer day included. Section II in Table 1 contains only the data for three teams instead of the original four, because one student withdrew from the course and his team consequently was disbanded; however, the remaining two students from that team continued in the course as non-team performers and their scores were used, in certain instances, as control data for evaluating the effects of team participation upon individual proficiency levels. While all of the split-half series and parallel team comparisons indicated that the second half of each team condition produced proficiency increments over the first half, only one comparison (Parallel A, B of Section I) produced an acceptable level of significance. In the series versus parallel team comparisons, Section I series team performances were superior to the parallel team's, while in Section II the series team performances were inferior to the parallel team performances. It should be noted that in the series versus parallel team comparisons the teams in Section I performed first under the series team condition and then under the parallel team condition. Teams in Section II performed under the parallel team condition first and the series team condition second. Thus, a potential time-position or sequence effect may have been operative and suggests that the team condition which occurred first, whether series type or parallel type, was the optimum one.

TABLE 1

ERROR RESPONSES OF SPLIT-HALF SERIES AND
PARALLEL TEAMS AND SERIES VERSUS
PARALLEL TEAMS

Teams	Section I*	Section II
Series A, B	204, 157 p (ns)[a]	227, 155 p (ns)[a]
Parallel A, B	297, 176 $p < .05$[a]	154, 143 p (ns)[a]
Series, Parallel	449, 534 p (ns)[b]	437, 382 p (ns)[b]

* Section I contained four teams, Section II contain-
ed three teams.

[a] Determined by t-tests.   [b] Determined by sign tests
for matched pairs (3).

The record of daily absences from class during
all team conditions revealed that Section I series
teams had no absences, Section I parallel teams had
35 percent, Section II series teams had 15 percent,
and Section II parallel teams had 50 percent of the
total number of absences. In summary, across sec-
tions, the series teams were responsible for 15 per-
cent of the total number of absences, while the par-
allel teams were responsible for 85 percent.

A further analysis of the split-half parallel teams
in relation to individual member's proficiency levels
showed that in both sections 72 percent of the origin-
ally low proficiency members improved their perfor-
mances over time (28% showed decrements); that is,
the majority of low proficiency members showed reg-
ular increments and made more contributions to the
team's output as time progressed. On the other hand,
43 percent of the initially high proficiency members
gave poorer performances (43% produced increments
and 14% showed no change) over time. Thus, there
appears to have been some sort of "leveling" factor
present wherein the initially high proficiency mem-
ber and the initially low proficiency member approach
a mutual or common proficiency level.

The foregoing result led us to construct a predic-
tion table concerning series-team and parallel-team
functioning and allowed us to show the effects of the
leadership position on team output. Specifically, in
addition to other information, Table 2 indicates the
relationship between individual proficiency and team
proficiency when each member occupies the role of
leader in a parallel team. Following an earlier anal-
ysis (2) with series teams, it was assumed that the
performance of team members was independent of
the team performance, and was a multiplicative func-
tion of the probability that each team member would
perform correctly on any one question on a quiz.
Moreover, in a parallel team situation, it was pre-
dicted that when a leader originally performs at low
proficiency levels, the addition of two redundant high
proficiency members would add substantially to team

output. On the other hand, when a parallel team leader
originally performed at high proficiency levels, it
was predicted that the addition of two redundant low
proficiency members would subtract substantially
from the team output. Proficiency values determin-
ed by the actual classroom performance levels of
individuals and teams are also shown in Table 2. For
the parallel teams, 86 percent of the team leaders
led their teams to actual proficiency levels above the
predicted range, while only 14 percent of the team
leaders placed their teams at actual proficiency lev-
els within, or below, the predicted range. In the
case of series teams and the predicted versus actu-
al proficiency levels, without exception the actual
levels were above the predicted levels of proficien-
cy for all teams. Thus, there were substantial team
gains under both series and parallel team conditions
and neither the multiplicative law for series teams,
nor the formula of decrements,

$$/M_1 \, ((L + M_2 \, (1.00 - L) \,))/$$

for parallel teams seemed to hold completely in the
present study which involved relatively complex
learning tasks.

Table 3 shows individual student proficiency lev-
els before and after holding membership on teams
under both series and parallel team conditions. The
results in Table 3 are reported in terms of gains of
results in Table 3 are reported in terms of gains of
the proportion of correct responses for both the
quizzes and the final exams which were given after
team formation and before team participation. Only
one student failed to show a gain on the quiz mate-
rial from the pre-team to post-team testing situa-
tions; on the same final exam, however, the student
did show a gain in proficiency. A calculation of the
average gain from pre-team to post-team conditions
for all students in the class who served on teams
showed a proportion gain of .20 for the quizzes and
.47 for the final exam. The scores from the two
students whose team disbanded early in the second
week showed average non-teams gains of only .10 for
the quizzes and .36 for the final exam. By virtue of
its small size, this 2-student control group data was
considered to be weak and no further statistical eval-
uations were made using it as non-treatment criteria.
All the students were originally part of a larger pop-
ulation which registered for the course; students were
randomly assigned to the treatment group by taking
the names on the lower half of a head-count sign-up
sheet which was made during the pre-course regis-
tration period. Both groups of students took the
course concurrently at the same time each day, one
group under the traditional lecture method and the
other under the 3-man team method. The final ex-
ams from both groups were compared in instances
where the same questions were given on both exams.
There were twelve common questions (out of a total
of seventy on each exam) using this procedure, and
it provided another basis for comparison of our group
with an untreated group. A t-test of the difference
between the error responses of the two groups show-
ed that the untreated group was far inferior to the
treatment group on the final exam questions ($t = 4.8$;
$df = 44$; $p < .001$) and again indicated the effective-
ness of the multiple quiz procedure used with the 3-
man teams.

The results from the pre-team and post-team

TABLE 2

PREDICTED AND ACTUAL INDIVIDUAL AND TEAM PROFICIENCY LEVELS*

Team		Member A	Member B	Member C	Predicted Series Team Level[a]	Actual Series Team Level	Predicted Parallel Team Level Under Leadership of Member[b]			Actual Parallel Team Level Under Leadership of Member		
							A	B	C	A	B	C
1	S	.86	.95	.70	.57	.84						
	P	.74	.88	.68			.66-.80	.66-.71	.71-.80	.98	.91	.79
2	S	.78	.89	.74	.51	.80						
	P	.77	.84	.77			.74-.82	.73-.74	.74-.82	.94	.87	.83
3	S	.83	.83	.82	.56	.83						
	P	.83	.85	.77			.75-.82	.75-.79	.79-.82	.94	.84	.82
4	S	.87	.64	.88	.48	.79						
	P	.75	.70	.77			.66-.71	.69-.71	.66-.69	.76	.92	.80
5	S	.74	.62	.69	.32	.69						
	P	.78	.64	.68			.60-.63	.63-.69	.60-.69	.80	.79	.85
6	S	.86	.90	.87	.67	.88						
	P	.91	.94	.92			.91-.93	.90-.91	.90-.93	.99	.94	.93
7	S	.79	.67	.69	.37	.72						
	P	.88	.65	.72			.62-.69	.69-.79	.62-.79	.95	.88	.85

* Table entries indicate proportion of correct responses.

[a] For Series (S) teams, predicted overall proficiency is computed using the multiplicative law of probability.

[b] For Parallel (P) teams, predicted overall proficiency is computed using the formula: $M_1 ((L + M_2 (1.00 - L)))$ where L denotes leader, M_1 and M_2 denote redundant members.

administration of the CTA and SSHA showed that on the second testing of the CTA, 70 percent of the students scored higher than in the first testing session (13% scored lower, and 17% showed no change); on the second testing of the SSHA, 44 percent of the students scored higher than in the first testing period (47% scored lower, and 9% showed no change in scores).

SUMMARY AND CONCLUSIONS

The model which we adopted (2) for describing information processing activities within 3-man series and 3-man parallel teams had only limited confirmation from actual classroom data. According to the model, in a series team all members had to perform efficiently for the team to have a high score on a quiz and all individual performances were "added" (i. e., the team contained non-redundant membership roles); in a parallel team only the leader had the final word or action as to what the team product on the quiz would be and some individual performances were "subtracted" (i. e., the team contained redundant membership roles). However, it was shown that predictions based upon the model were supported fully only in certain circumstances involving increments in individual member proficiency levels (see Table 3), less completely in circumstances involving proficiency increments and decrements in split-half series and parallel teams and in series versus parallel teams (see Table 1), and not at all in circumstances involving the role of leader upon team proficiency (see Table 2). Specifically, the six hypotheses which were set up to be tested in an authentic classroom situation had the following degrees of confirmation. The notion in Hypothesis 1 concerning series team increments over time was supported somewhat, but the notion of series team superiority over the parallel team structure regarding individual proficiency levels was not supported. Hypotheses 2, 3, and 4 were not supported by the data. Hypothesis 5 concerning increments in

TABLE 3

INDIVIDUAL PROFICIENCY LEVELS BEFORE AND AFTER MEMBERSHIP ON SERIES AND PARALLEL TEAMS*

| Student [a] | Performance on Quizzes Before Membership | Performance on Quizzes After Membership | Gain | Performance on Final Exam | | Gain |
				First Testing	Second Testing	
1	.43	.89	.46	.33	.85	.52
2	.85	.95	.10	.27	.88	.61
3	.68	.76	.08	.27	.72	.45
4	.56	.89	.33	.37	.86	.49
5	.80	.98	.18	.35	.88	.53
6	.52	.82	.30	.37	.69	.32
7	.72	.83	.11	.30	.68	.38
8	.75	.90	.15	.43	.89	.46
9	.55	.79	.24	.35	.74	.39
10	.73	.90	.17	.45	.86	.41
11	.53	.70	.17	.34	.57	.23
12	.73	.89	.16	.17	.89	.72
13	.68	.86	.18	.35	.70	.35
14	.49	.60	.11	.40	.60	.20
15	.86	.95	.09	.30	.83	.53
16	.55	.94	.39	.24	.96	.72
17	.65	.98	.33	.38	.99	.61
18	.80	.95	.15	.26	.96	.70
19	.52	.83	.31	.17	.75	.58
20	.72	.93	.21	.36	.70	.34
21	.75	.70	None	.38	.65	.27

* Table entries indicate proportion of correct responses.

[a] The data of two students from the disbanded team are not included.

individual proficiency from the beginning of the course to the end of the course was supported by the data from the multiple quizzes and exams. Hypothesis 6 was partially supported when a majority of students showed an increase in their scores on a critical thinking test but not on a study-habits survey.

In conclusion, the partial failure of the present study to verify hypotheses which were derived from previous experimental studies on team performance should not detract from the fact than an extrapolation of reinforcement theory from the laboratory to the classroom situation was made successfully and the method used to test the hypotheses was effective in raising individual student's proficiency levels. For instance, while our study did not follow an earlier analysis (2) of the operator of leadership position in series versus parallel teams, members of both types of teams in our course made substantial gains on the assigned material. Thus, the primary educational and behavioral objectives that were set up were achieved in spite of the incomplete success concerning the secondary objective of testing hypotheses. We are currently employing the team-testing

technique as part of our regular semester psychology course which meets for three sessions a w e e k; the first session is the instructor's lecture; the second is student recitation and discussion; and the third is an examination period which employs the series and parallel team structures.

FOOTNOTES

1. The present study was made possible through the assistance and cooperation of Dr. John Riggs, Dr. Ray Cattani, Dr. James Scoresby, a n d Mr. Jerrell Ferguson.

2. Instructions which were given to the students concerning series and parallel team functioning and the "rules" for processing the quizzes in the teams tended to be lengthy and are omitted. You can receive copies of these instructions by writing to Dr. Jon E. Roeckelein, Department of Psychology, Arizona State University, Tempe, Arizona 85281.

REFERENCES

1. Campbell, D. T. ; Stanley, J. C. , "Experimental and Quasi-experimental Designs for Research on Teaching," in Gage, N. L. (ed.), Handbook of Research on Teaching, Rand McNally, Chicago, Illinois, 1962, pp. 171-246.

2. Glaser, R. ; Klaus, D. J. , "A Reinforcement Analysis of Group Performance," Psychological Monographs: General and Applied, 80:(no. 13), 80:(no. 621), 1966.

3. Hays, W. L. , Statistics for Psychologists, Holt, Rinehart, and Winston, New York, 1963.

TEST STATISTICS AS A FUNCTION

OF ITEM ARRANGEMENT

DAVID M. SHOEMAKER[1]
Oklahoma State University

ABSTRACT

Using a technique that controlled exposure of items, the investigator examined the effect on mean test score, item difficulty index, and reliability and validity coefficients of the reordering of items within a power test containing ten letter-series-completion items. The results suggest that effects on test statistics from item rearrangement are, generally, minimal. The implication of these findings for test designs involving an item sampling procedure is that performance on an item is minimally influenced by the context in which it occurs.

ITEM SAMPLING, a procedure involving the planned confounding of subjects and test items, has been demonstrated empirically (4, 5, 8) and algebraically (7) to be a valuable experimental design in test research. However, one critical assumption in item sampling is that performance on an item is minimally influenced by the context in which it occurs. This assumption has ostensibly been evaluated in several investigations (e. g., 1, 6, 9) which have directly or indirectly examined the effect on test statistics of rearranging items within a test. The standard procedure for manipulating item sequence has been that of reordering items on the printed page. The assumptions are (a) an examinee responds to items in the order in which they appear on the printed page, and (b) after responding to an item, an examinee does not give it additional consideration. For power tests, both assumptions are questionable and suggest that the traditional procedure should be modified to eliminate uncontrolled item review on the part of the examinee. The point to be made is that any investigation attempting manipulation of test item arrangement must incorporate a procedure which strictly controls the exposure of test items to examinees.

An additional consideration in examining test statistics as a function of item arrangement is the content of the items. It is hypothesized that item rearrangement will significantly influence test statistics for those content areas where an item solution is contingent upon the generation of concepts or algorithms. Contrast, for example, a set of vocabulary items with a set of letter-series-completion items (sample item: DABEDCFEFG ?) of the type found in the Thurstone Letter Series Completion Test. Determining the appropriate solutions for vocabulary items does not appear, at least intuitively, to depend upon the generation of algorithms; however, results obtained by Simon and Kotovsky (10) on the acquisition of concepts for sequential patterns strongly suggest that the opposite is true with letter-series-completion items. As algorithms vary in complexity, it may be argued that experience with less complex algorithms will facilitate the generation of more complex algorithms.

Using a technique which controlled item exposure to examinees, the present investigator examined the effect on mean test score, item difficulty index, reliability, and validity coefficients of the reordering of items within a test. The items selected for consideration were of the letter-series-completion type.

METHOD

From results obtained in a series of item analyses of letter-series-completion items, a set of ten items (referred to hereafter as the Letter Series Test) was selected. Difficulty indices for the items were rectangularly distributed with a range of .167 to .917. Four experimental forms of the Letter Series Test were constructed. In Forms 1, 2, and 3, the items were arranged within the test booklet with one item per page; in Form 4, all items were on one page. In Form 1, the items were sequenced in order of ascending difficulty; Form 2, in descending difficulty; Form 3, randomly sequenced; and Form

4, same sequence as Form 3. After a brief introduction involving five practice items, the instructions were as follows:

> There are ten items in the test—one item on each page. Work the items in the order in which they appear in the test. After working an item, fold that page behind the test and proceed to the next item. Do not return to any item after you have once worked it or have tried to work it. If you are unable to work a particular item, fold that page behind the test and proceed to the next item. If any answer to an item would just be a "wild guess" on your part, skip that item and go on to the next one. There is no time limit on the test.

Identical sample items were used in the four forms of the Letter Series Test. For Form 4, however, the instructions were modified to exclude reference to items being printed on separate pages.

Four classes of college students from introductory courses in psychology, sociology, and educational psychology were selected as Ss. In three of the four classes Forms 1, 2, and 3 were distributed among students in an alternating fashion, in the fourth class only Form 4 was administered. (The undesirable confounding of Form 4 with a specific group of examinees was necessitated by the nature of the instructions on the test booklet.) A 20-item Number Series Test (sample item: 2 4 6 8 11 13 15 ?) having a 6-minute time limit was also administered to each examinee. Administering the Number Series Test to all examinees permitted an additional technique of data analysis, namely, analysis of Letter Series tests results by subgroups homogeneous (above median, below median) on Number Series Test. While both tests were administered to all examinees, the order of test administration was counterbalanced across classes.

No unusual circumstances arose during the administration of any test. Examinees appeared to follow the instructions as outlined in the test booklet.

RESULTS

Several analyses of variance (ANOVA) were performed on the Number Series and Letter Series test scores and, as the significance levels of the computed F statistics were generally greater than 15 percent, the ANOVA results can be briefly summarized as follows: (a) For the Number Series Test scores, no significant differences were observed between classes, between positions of administration in the test battery, or between sexes. Identical analyses involving Letter Series Test scores produced similar results. (b) Using proportional cell frequencies, a 4x2 analysis of variance of Letter Series Test scores was performed involving experimental test forms and groups (above median, below median on Number Series Test). Both the forms ($F_{3,124} = 1.684$, $.25 > p > .15$) and the forms X group interaction ($F_{3,124} = .953$) effects were judged to be nonsignificant. The group effect ($F_{1,124} = 9.523$, $.005 > p > .001$) was judged to be significant and considered a confirmation of the method of grouping examinees into above-median and below-median groups on the basis of the Number Series Test score.

An item difficulty index (proportion answering the item correctly) was computed for each item in each experimental test form for (a) the total sample, (b) each sex, and (c) above-median examinees and below-median examinees on the Number Series Test. Mean test score, standard deviation of test scores, and Kuder-Richardson Formula 20 reliability coefficients were computed for each experimental test form in each analysis. As similar results were obtained in all analyses, only the results for the total sample are given in Table 1. The inter-form intercorrelations among item difficulty indices for each experimental test form for the total sample are given in Table 2. The validity coefficients for each form of the Letter Series Test are given in Table 3. In addition, the regression equations for each experimental test form on Number Series Test scores were computed.

TABLE 1

ITEM DIFFICULTY INDICES, MEANS, AND STANDARD DEVIATIONS, AND KUDER-RICHARDSON FORMULA 20 RELIABILITY COEFFICIENTS PER EXPERIMENTAL TEST FORM FOR TOTAL SUBJECT SAMPLE

Item Number	Form 1	Form 2	Form 3	Form 4
1 (.917)[a]	.950	.975	.917	.913
2 (.792)	.875	.950	.972	.913
3 (.708)	.875	.625	.667	.826
4 (.583)	.650	.600	.500	.522
5 (.500)	.675	.700	.694	.696
6 (.417)	.650	.450	.500	.609
7 (.292)	.800	.675	.750	.957
8 (.250)	.350	.350	.417	.609
9 (.208)	.500	.475	.417	.652
10 (.167)	.650	.625	.500	.565
N	40	40	36	23
X	6.98	6.43	6.36	7.30
SD	2.13	1.99	1.90	1.88
KR20	.668	.573	.509	.557

[a]Computed difficulty index of item from item-analysis performed prior to the construction of 10-item experimental test forms.

The individual regression equations are as follows:

$$N = .742\, L_1 + 7.050$$

$$N = .417 \, L_2 + 8.871$$

$$N = .264 \, L_3 + 9.904$$

$$N = .735 \, L_4 + 6.284$$

Since each examinee has taken both Number and Letter Series tests, scores on the experimental test forms could be compared by means of the analysis of covariance procedure described by Gulliksen and Wilks (3).

TABLE 2

INTER-FORM PRODUCT-MOMENT CORRELATION COEFFICIENTS FOR ITEM DIFFICULTY INDICES PER EXPERIMENTAL TEST FORM FOR TOTAL SUBJECT SAMPLE

	Experimental Test Form			
	1	2	3	4
1		.857	.862	.777
2			.936	.723
3				.889
4				

TABLE 3

VALIDITY COEFFICIENTS: PEARSON PRODUCT-MOMENT CORRELATION COEFFICIENTS FOR EXPERIMENTAL FORMS OF LETTER SERIES TEST (L) WITH NUMBER SERIES TEST (N)

Letter Series Test	N	r_{NL}	Var (N)	Var (L)
Form 1	40	.534	8.724	4.537
Form 2	40	.405	4.248	3.960
Form 3	36	.213	5.576	3.610
Form 4	23	.527	6.836	3.534

The hypotheses of equal errors of estimate, equal regression line slopes, and equal intercepts were tested; the significance levels of the obtained differences were, in each case, greater than .30. The item difficulty indices per item position across experimental test form are given in Table 4.

DISCUSSION

The fact that no significant differences in Number Series Test scores were observed between classes suggested that treatment effects were not being confounded with classes. Data were then pooled across classes to increase the sample size per experimental test form and hence the stability of the computed

TABLE 4

ITEM DIFFICULTY INDICES PER ITEM POSITION IN LETTER SERIES TEST

Item Position	Total Sample (N = 139)	Males (N = 62)	Females (N = 77)
1	.748	.790	.714
2	.604	.613	.597
3	.568	.613	.532
4	.784	.774	.792
5	.532	.565	.506
6	.698	.694	.701
7	.633	.645	.623
8	.669	.694	.649
9	.770	.790	.753
10	.691	.677	.701

test statistics. Another possible confounding was position of experimental test in the test battery. However, for the Letter Series Test the mean test score for those examinees receiving the test first in the battery was not significantly different from the mean test score of those receiving the test in second position. A similar result was obtained for the Number Series Test. The possibility, then, that test performance on the second test in the battery was facilitated by taking the first test was not supported by the data. Furthermore, when item difficulty indices were computed for each item position across experimental test forms, items occurring toward the end of a test were not answered correctly more often than items appearing at the beginning of a test. The effect on mean test score of item rearrangement was found to be minimal; however, the trend consistently found in all analyses was that of Form 4 having the largest mean test score followed consecutively by Forms 1, 2, and 3. Form 4, it will be recalled, was the experimental test form with all items printed on one page; thus, it was possible for an examinee to modify a previous answer based on any insights acquired in taking the test.

One noticeable effect regarding items was the change in item difficulty indices between the pilot study and the principal investigation. Items found to be most difficult in the pilot study were generally not as difficult in the main investigation. It should be noted that this finding, while interesting and undoubtedly a reflection of sampling bias, does not in any way qualify the results obtained in this study. The item difficulty indices in Table 1, specifically those computed for items 3, 6, 7, 8, and 9, suggest that, when item exposure is strictly controlled, the item difficulty index for certain items may be a function of the position of that item within the test.

The results given in Table 1 suggest that the effect on the Kuder-Richardson Formula 20 reliability coefficient resulting from item rearrangement within a test is slight. Similar results were reported by Brenner (1) for the Kuder-Richardson Formula 8 coefficient and by Flaugher and others (2) for the Kuder-Richardson Formula 20 coefficient.

Perhaps the most striking effect on test statistics resulting from item rearrangement was the change in validity coefficients. Flaugher and others (2) reported changes in validity coefficients of magnitude .01 to .03 when items were reordered within a test; however, the differences observed in the present investigation were considerably greater. Furthermore, the relative differences among validity coefficients do not appear to be attributable to the magnitudes of the variances used in computing the correlation coefficients.

The differences among regression equations were found not to be significant. Nonetheless, if these regression equations were to be used indiscriminately in a testing program, an examinee having, for example, a score of 2 on the Letter Series Test upon receiving Form 3 would have approximately a 2.5 point advantage in Number Series Test score over an examinee of equal ability who received Form 4. The practical implication of this point might well be questioned. Two and one-half points fall well within the range of possible error from other sources, and therefore, might be discounted. However, this is additional error beyond that arising from other sources.

FOOTNOTE

1. Dr. Shoemaker's current address is: Southwest Regional Laboratory for Educational Research and Development, 11300 La Cienega Blvd., Inglewood, California 90304.

REFERENCES

1. Brenner, M. H., "Test Difficulty, Reliability, and Discrimination as Functions of Item Difficulty Order," Journal of Applied Psychology, 48:98-100, 1964.

2. Flaugher, R. L., Melton, R. S.; Myers, C. T., "Item Rearrangement Under Typical Test Conditions," Educational and Psychological Measurement, 28:813-824, 1968.

3. Gulliksen, H.; Wilks, S. S., "Regression Tests for Several Samples," Psychometrika, 15:91-114, 1950.

4. Johnson, M. C.; Lord, F. M., "An Empirical Study of the Stability of a Group Mean in Relation to the Distribution of Test Items Among Students," Educational and Psychological Measurement, 18:325-329, 1958.

5. Lord, F. M., "Estimating Norms by Item-Sampling," Educational and Psychological Measurement, 22:259-269, 1962.

6. Mollenkopf, W. G., "An Experimental Study of the Effects on Item-Analysis Data of Changing Item Placement and Test Limit," Psychometrika, 15:291-315, 1950.

7. Osburn, H. G., "A Note on Design of Test Experiments," Educational and Psychological Measurement, 27:797-802, 1967.

8. Plumlee, L. B., "Estimating Means and Standard Deviations from Partial Data—An Empirical Check on Lord's Item Sampling Technique," Educational and Psychological Measurement, 24:623-631, 1964.

9. Sax, G; Carr, A., "The Effects of Various Forms of Item Arrangements on Test Performance," Journal of Educational Measurement, 3:309-311, 1966.

10. Simon, H. A.; Kotovsky, K., "Human Acquisition of Concepts for Sequential Patterns," Psychological Review, 70:534-546, 1963.

THE COHORT-SURVIVAL RATIO METHOD

IN THE PROJECTION OF SCHOOL ATTENDANCE

WILLIAM J. WEBSTER
American Institutes for Research

ABSTRACT

The purpose of this study was to test the hypothesis that educational attendance projection problems could be formulated in such a manner that the regression model could be used in the analysis thereof. Two projection approaches, the Cohort-survival ratio approach and a comparable regression approach, were compared using a sample of twenty-five Michigan school districts. The regression approach provided significantly better estimates than did the Cohort-survival ratio approach.

THE ACUTE need for educational planning has long been realized. One essential facet of this planning involves the estimation of numbers of pupils to be housed at relevant future dates. Accurate enrollment projections would greatly assist educators in making important decisions affecting future educational and fiscal allocations. However, despite the numerous and important uses of accurate enrollment projections, the formulae currently most used for projection purposes display a notable lack of desired precision (4, 5).

The purpose of the present study was to test the hypothesis that school attendance projection problems can be formulated in such a manner that the regression model can be used to obtain more accurate estimates of future public school enrollment than can be obtained from the most popular of the ratio projection methods currently used by educators. The distinction between ratio and regression analysis methods is made on the basis of the measure of relationship used in the computation of projected enrollments. Ratio methods use the ratio or proportion of a given predictor to the past, while regression methods depend upon coefficients of correlation and multiple correlation. Such statistics purport to measure degree of association, and provide projections based on a trend that has been established in the past. It is believed that this procedure represents a new and unique method of estimating future public school educational attendance patterns, although similar procedures have occasionally been experimented with in higher education (1).

METHODOLOGY

The most popular as well as the most accurate of the enrollment projection techniques currently used by educators for projecting future public school enrollment is the Cohort-survival ratio method. This method depends upon the relationship of birth-to-grade and grade-to-grade statistics for a number of years. Predictor variables are births and past tendencies of school children to advance from one grade to the next. The Cohort-survival ratio method was one of the projection methods examined in this study.

A regression analysis, utilizing the same predictor variables as the Cohort-survival ratio method, was also employed.

Cohort-survival ratio analysis and regression analysis were used to develop projections of elementary (K-8) and secondary (9-12) enrollment for a sample of twenty-five school districts for the years 1965 and 1968. Projections were based on data collected from the years 1955-1960. Actual enrollment during the years 1965 and 1968 was used as the criterion against which to judge the effectiveness of each of the two projection methodologies. The procedures

used in obtaining projections from the two models are outlined below.

1. The actual number of births by place of residence by year for the period 1950-1960 were obtained for each district in the sample.

2. The actual enrollment figures by grade for kindergarten through twelfth grade for the years 1955-1960 were obtained for each district in the sample.

3. The mean number of births by place of residence for the years 1952-1960, 1948-1951, 1955-1963, and 1951-1954, was computed for each district in the sample. These dates correspond to the birth dates of children in elementary school in 1965, in secondary school in 1965, in elementary school in 1968, and in secondary school in 1968, respectively.

4. For the Cohort-survival ratio method, the average ratio of survivors from birth, 5 years earlier, to kindergarten; and from grade to successive grade, was computed for each district in the sample. Ratios were based on data corresponding to the period 1955-1960. This procedure resulted in thirteen separate retention ratios, one corresponding to the average percentage of students "surviving" the transition into each successive grade, kindergarten through twelve. According to the nature of desired projections (elementary or secondary enrollment) and the year to which they were to be made, one of the means referred to in step three was chosen as a starting point. (For example, the mean number of births by place of residence for the years 1952-1960 was used as a starting point in projecting elementary enrollment to 1965.) The average retention ratio of kindergarten enrollment to births 5 years earlier was multiplied by the relevant mean number of births to obtain a projection of kindergarten enrollment in the desired year. This estimate was then multiplied by the average retention ratio of kindergarten to first grade to obtain a projection of first grade enrollment in the desired year. This procedure was repeated seven additional times; 1 to 2, 2 to 3, 3 to 4, 4 to 5, 5 to 6, 6 to 7, and 7 to 8. Each time different input variables were used to obtain projections of enrollment in each of the nine elementary grades (K-8).

If secondary enrollment was being projected, the same procedure was used to obtain an estimate of eighth grade enrollment. Starting with the appropriate mean number of births by place of residence, nine retention ratios, one corresponding to transition into each of the grades, K-8, were successively progressed through to obtain the required eighth grade projection. Using that projection as a base, enrollment was projected to ninth grade. This procedure was repeated three additional times; 9 to 10, 10 to 11, and 11 to 12. Each time different input variables were used to obtain projections of enrollment in each of the four secondary grades (9-12).

The separate projections were finally summed over grades to obtain the relevant projections of elementary (K-8) and secondary (9-12) enrollment.

In summary, for each grade:

$$E_n = E_{n-1} \ (R_{n-1 \text{ to } n})$$

Where

E_n = Grade n enrollment in year x.

E_{n-1} = Grade n-1 enrollment in year x.

$R_{n-1 \text{ to } n}$ = The average retention ratio denoting the average survival rate of students from grade n-1 to grade n for the period 1955-1960.

Note that mean numbers of births by place of residence were used as the basis for projections because only estimates of elementary and secondary enrollment as entities were desired. Grade level projections were treated only as components of the entire elementary or secondary projection, rather than as entities within themselves. A much more complex procedure would be required to obtain relatively precise projections of enrollment by grade level.

5. For the regression approach, the same predictor and criterion variables were used. Instead of averaging ratios, the least-squares criterion was used to obtain thirteen equations of the form Y = a+ bx, that, is one equation for transition into each grade—kindergarten through twelve. Once these equations were obtained, all procedures were identical to those outlined above.

In summary, for each grade:

$$E_n = a + b \ (E_{n-1})$$

Where:

E_n = Grade n enrollment in year x.

a = The Y intercept of the line defined by the relationship between the predictor and criterion variables for the period 1955-1960,

b = The slope of the line defined by the relationship between the predictor and criterion variables for the period 1955-1960,

E_{n-1} = Grade n-1 enrollment in year x.

The population utilized in this study consisted of every public school district in the State of Michigan that had grades kindergarten through twelve and was largely coterminous with a city of 10,000 population or more (a total of some eighty districts). Since it was recognized that districts differ in terms of growth characteristics, the population was divided into five strata on the basis of the following decision rules involving past population growth characteristics of the principal city in each district.

1. School districts classified in Stratum A are located in communities that increased by more than 100 percent during the decade 1950-1960, in both general population and in the number of households.

2. School districts classified in Stratum B are located in communities that increased by 50 percent or more in either or both general population and number of households, but 100 percent or less in one of the classification variables, during the decade 1950-1960.

3. School districts classified in Stratum C are located in communities that have experienced an increase of 10 percent or more, but less than 50 percent in both general population and number of households during the decade 1950-1960.

4. School districts classified in Stratum D are located in communities that have experienced an increase of less than 10 percent in both general population and number of households, or, an increase of less than 10 percent in one of the classification variables, and an increase of 10 percent or more in the other, during the decade 1950-1960.

5. School districts classified in Stratum E are located in communities that have experienced a decrease in either or both general population and number of households during the decade 1950-1960.

The final sample consisted of twenty-five school districts, five drawn randomly from each of the previously described strata.

Goodness of fit between estimated and actual school district enrollment in 1965 and 1968 formed the basis for evaluation. Each of the twenty-five districts that comprised the sample was characterized by eight separate estimates of different aspects of its enrollment, including projections of elementary enrollment in 1965, secondary enrollment in 1965, elementary enrollment in 1968, and secondary enrollment in 1968, by Cohort-survival ratio analysis and by regression analysis. Goodness of fit was determined by subtracting the actual enrollment figure from the projected enrollment figure for each of the aforementioned estimates, dividing each by its actual enrollment figure, and reporting the resulting statistic, the W coefficient, a measure of the percentage of error in each projection. The smaller the W coefficient, the better the estimate.

In order to gain insight into the relative "goodness" of the two projection approaches, W coefficients were ranked within each of the four estimates by district. By examining these ranks it is possible to make some inferences concerning the relative efficiency of the two projection methods for estimating different aspects of public school enrollment within districts and strata. In addition, W coefficients were summed across the four sub-category estimates within districts to obtain composites. These composite W coefficients were also ranked.

A simple measure of the percentage of error present in each projection, termed the W coefficient by the author, was chosen over a chi-square comparison of frequencies for a number of reasons. First, a simple descriptive statistic was all that was needed to describe the relative goodness of the two approaches. There was no interest at this point in determining whether or not a given projection differed significantly from a population value, or, whether the formulae performed differentially within strata. Due to the unique aspects of the projection problem, in the first case we would have been placed in the undesirable position of wanting to fail to reject a null hypothesis, and in the second case we would have been faced with an extremely small N. Second, chi-square requires that observations be independent both within and between groups. Though projections are certainly independent between districts, the same equations were used within districts to obtain more than one projection. For this reason within district projections were lumped together through a composite W coefficient and ranked, thus eliminating the problem of lack of independence within districts while maintaining the independence between districts. Finally, the W coefficient is more easily interpretable in the context used here, as it presents simple percentage of error while preserving the direction of deviation, information that is lost with chi-square.

Since only two projection methodologies were examined in this study, the only possible ranks for each projection within each district are a rank of 1 or 2. All possible independent observations can thus be conceived of as falling into either one or the other of two discrete classifications. Since it had been hypothesized that regression analysis would yield more accurate estimates of future public school enrollment than would Cohort-survival ratio analysis, the important question became one of whether the probability associated with the number of "1" ranks received by regression analysis was greater than could be expected by chance. The Binomial Test was used to test the hypothesis.

QUALIFICATIONS INHERENT IN THE STATISTICAL PROJECTION METHODS UTILIZED

The major qualification that must be stated for both Cohort-survival ratio analysis and regression analysis involves the basic tenet that the environment in the period for which projections are made must remain similar to that of the period from which data were drawn. To the extent that this assumption is violated, enrollment projections will be in error.

Regression analysis requires several additional assumptions. The most important of these states that the line of best fit which specifies the relationship between the predictor and criterion variables be a straight line. To the extent that the data deviate from linearity, projections derived from the linear model will be in error. This assumption is not seen as a problem in the present study since most functions can be best approximated by a straight line over a short period of time or small number of data points.

A second basic assumption underlying the application of regression methods is that the error terms in the regression model are independent. Cochrane and Orcutt have demonstrated that the usual application of standard least-squares methodology to relationships containing high positively auto-correlated error terms, while producing unbiased estimates, results in a marked decline of the variances of both the correlation coefficient and the regression coefficient as the error terms become more random (2). This phenomenon results in a serious underestimate of true error variance in cases where a high degree of auto-correlation is present. Given the fact that there is a high probability of auto-correlated error terms in time-series data, it is doubtful that confidence intervals computed around the present projections in the traditional least-squares manner would add significantly to the available data. Since the projections themselves are not biased by the existence of auto-correlated error terms, the problem is reserved for future study. Readers interested in the problem of constructing meaningful confidence

limits around projections made from data character-
ized by auto-correlated error terms are referred to
the discussion of this problem presented by Johnson
(4).

Other assumptions of the regression model, name-
ly those of normality and common variance, are of
little consequence since moderate variations in those
assumptions have little effect on the results obtained
from the model.

RESULTS

The results of the study are presented in Tables
1 and 2. The information contained in the tables in-
cludes, for the Cohort-survival ratio method (Table
1) and regression analysis (Table 2), by district,
respectively:

1. The code letter for each district in the
   sample;

2. Projected Elementary Enrollment, 1965
   (EE-1965);

3. The W coefficients for the 1965 elemen-
   tary projections (W);

4. The rank of the W coefficients of the 1965
   elementary projections relative to the projec-
   tions yielded by the other method (R = 1 or 2);

5. Projected Secondary Enrollment, 1965
   (SE-1965);

6. The W coefficients for the 1965 secondary
   projections (W);

7. The rank of the W coefficients of the 1965
   secondary projections relative to the projec-
   tions yielded by the other method (R=1 or 2);

8. Projected Elementary Enrollment, 1968
   (EE-1968);

9. The W coefficients of the 1968 elementary
   projections (W);

10. The rank of the W coefficients of the 1968
    elementary projections relative to the pro-
    jections yielded by the other method (R = 1
    or 2);

11. Projected Secondary Enrollment, 1968
    (SE-1968);

12. The W coefficients for the 1968 secondary
    projections (W);

13. The rank of the W coefficients of the 1968
    secondary projections relative to the projec-
    tions yielded by the other method (R = 1 or 2);

14. The rank of the composite W coefficients
    (R). The W coefficients are summed within
    districts and projection methodologies across
    the four separate enrollment estimates. The
    resulting composite W coefficient is then rank-
    ed across the two projection methodologies
    (R = 1 or 2) and provides the basis for assess-

ing the relative performance of the two meth-
odologies.

If the two projection methodologies were perform-
ing alike, the probability of obtaining a rank of 1 on
any given trial should be .50 for either Cohort-sur-
vival ratio analysis or regression analysis. There-
fore, H_0: $P_1 = P_2 = \frac{1}{2}$. Since it had been predicted
that regression analysis would produce more accu-
rate projections, H_1: $P_1 > P_2$. It will be seen from
Table 2 that regression analysis received eighteen
"1" ranks out of a total of twenty-five. The α level
associated with such an occurrence under the bino-
mial distribution is .022; therefore, since .022 <
.05, H_0 is rejected. The data therefore suggest that
regression analysis provides more accurate projec-
tions of future public school enrollment than does
Cohort-survival ratio analysis.

SUMMARY

The results of the study suggest that regression
analysis provides a viable approach to the projection
of future public school attendance. Standard com-
puter programs are readily available to provide the
least-squares projections required. It is recom-
mended that individual school districts, as well as
state departments of education, explore the feasi-
bility of utilizing regression analysis for projecting
future public school enrollment.

DISCUSSION

The general superiority of the regression ap-
proach can be explained in terms of the previous
discussion of the differences in measure of relation-
ship used in the computation of the two types of pro-
jection equations. By averaging ratios, the Cohort-
survival ratio method washes out year-to-year
differences in variability, thus contributing to error
in the projections. In the case of regression analy-
sis, on the other hand, the same differences in var-
iability merely lower the size of the correlation co-
efficients. The fact that projections yielded by the
Cohort-survival ratio method were relatively more
accurate for districts classified within Strata C and
D, districts that are characterized by the smallest
amount of year-to-year differences in variability of
any of the districts considered in this study, would
seem to support this contention.

Regression analysis was not always superior be-
cause of the nature of the projection problem. Due
to the fact that no observations were recorded be-
tween the last year on which the projection equations
were based (1960) and the year to which projections
were to be made (1965 or 1968), there was no us-
able information for the intervening period of time.
Therefore, if the assumption of stable conditions, or
at the very least regular rates of change in the var-
iables of interest, was violated within a given dis-
trict, it became a matter of chance as to which pro-
jection equation provided the best estimates. That
there was no consistent relationship between the size
of the multiples, all of which were > .85, and wheth-
er or not regression analysis received a rank of
"1" within a given district, suggests
that this was the case.

PROJECTIONS BY THE COHORT-SURVIVAL RATIO METHOD

District	EE-1965	W	R	SE-1965	W	R	EE-1968	W	R	SE-1968	W	R
A	4836	+.0447	2	2462	+.1531	2	4917	+.0513	2	2580	+.1212	2
B	11108	+.0865	2	2800	-.0607	1	11796	+.0880	2	3317	+.0617	2
C	3684	+.0083	1	1086	-.0473	1	3697	+.0027	2	1190	-.0333	1
D	24122	+.0543	2	6225	-.1131	1	28126	+.1182	2	8047	-.1901	2
E	7861	+.5353	2	1970	+.0071	2	8023	+.6087	1	2067	-.0203	2
F	13808	+.1557	2	5702	+.4692	2	16508	+.1758	2	6086	+.2823	2
G	12168	+.0615	2	4360	-.0432	2	11484	-.0334	2	5348	-.0505	1
H	10242	+.1142	2	4218	+.0638	2	10751	+.1308	2	4503	+.1082	2
I	9662	+.1585	2	3605	+.2138	2	10444	+.1975	2	4115	+.2117	2
J	14746	+.0621	2	4728	-.1676	2	13390	+.0063	1	6411	+.0638	2
K	5995	+.0805	2	1849	+.0063	1	6519	+.0936	2	2204	-.0230	1
L	15022	+.1232	2	7149	-.1484	2	13728	+.0329	2	8439	-.0621	2
M	16771	+.2882	2	4758	+.0016	1	15768	+.1921	2	5196	+.0220	1
N	16569	-.0175	1	5449	-.0144	1	16692	-.0525	1	5896	+.0169	2
O	6225	+.0738	2	2283	-.1401	2	5628	-.0164	2	2998	+.1296	2
P	8117	-.0793	2	4688	-.0988	2	8840	-.0246	1	4820	-.1159	2
Q	23460	-.0178	1	8780	+.1236	2	22779	-.0606	2	9806	+.1176	2
R	5307	-.0502	2	2119	-.1396	1	5257	-.0495	2	2256	-.1075	2
S	3447	-.0653	1	2392	+.1267	2	3376	-.4433	1	2339	-.0484	1
T	15463	-.0040	2	6103	-.0073	1	14918	-.0722	1	7290	-.1447	1

Table 1 is continued on following page.

TABLE 1 (Continued from preceding page)

District	EE-1965	W	R	SE-1965	W	R	EE-1968	W	R	SE-1968	W	R	R
U	7685	-.0808	1	3166	+.0703	1	7777	-.0411	2	3192	+.0813	1	1
V	2375	.0000	1	969	-.2376	2	2085	-.0552	1	926	-.0776	2	2
W	4150	-.0253	2	2006	+.0979	2	3718	-.1019	2	2079	+.0038	1	2
X	1778	-.0787	2	760	+.0160	2	1666	-.1304	2	756	-.0307	1	2
Y	6069	-.0512	2	1890	-.2984	2	5615	-.0832	2	2829	+.0540	2	2

TABLE 2

PROJECTIONS BY REGRESSION ANALYSIS

District	EE-1965	W	R	SE-1965	W	R	EE-1968	W	R	SE-1968	W	R
A	4719	+.0194	1	2237	+.0477	1	4782	+.0224	1	2346	+.0195	1
B	10567	+.0336	1	2968	-.0949	2	10967	+.0116	1	2976	-.0473	1
C	3614	-.0114	2	1020	-.1052	2	3680	-.0018	1	1167	-.0519	2
D	23933	+.0465	1	7857	+.1193	2	27851	+.1073	1	10008	+.0072	1
E	7619	+.4880	1	1957	+.0005	1	8117	+.6276	2	2189	+.0374	2
F	11626	-.0268	1	3624	-.0662	1	13348	-.0492	1	4085	-.1392	1
G	11576	+.0099	1	4650	+.0204	1	11533	-.0293	1	4684	-.1684	2
H	9869	+.0737	1	3781	-.0464	1	10413	+.0952	1	3942	-.0297	1
I	8266	-.0088	1	2659	-.1047	1	8538	-.0209	1	2897	-.1469	1
	13930	+.0033	1	5635	-.0079	1	...9	+.0251	2	5667	-.0595	1

TABLE 2 (Continued from preceding page)

District	EE-1965	W	R	SE-1965	W	R	EE-1968	W	R	SE-1968	W	R	R
K	5304	$-.0620$	1	1729	$-.0812$	2	5652	$-.0518$	1	1863	$-.1742$	2	2
L	13681	$+.0229$	1	8306	$-.0105$	1	13388	$+.0073$	1	8489	$-.0565$	1	1
M	13886	$+.0666$	1	4521	$-.0482$	2	13661	$+.0328$	1	4528	$-.1093$	2	1
N	15690	$-.0696$	2	5685	$+.0282$	2	15705	$-.1079$	2	5720	$-.0134$	1	2
O	5986	$+.0326$	1	2316	$-.1276$	1	5803	$+.0141$	1	2517	$-.0516$	1	1
P	8364	$-.0513$	1	5426	$+.0430$	1	8540	$-.0577$	2	5390	$-.0113$	1	1
Q	24356	$+.0196$	2	8347	$+.0682$	1	23207	$-.0430$	1	9766	$+.1130$	1	1
R	5629	$+.0073$	1	2092	$-.1506$	2	5535	$+.0007$	1	2557	$+.0114$	1	1
S	3339	$-.0946$	2	2078	$-.0211$	1	3349	$-.4511$	2	2030	$-.1741$	2	2
T	14278	$-.0803$	1	6332	$+.0299$	2	14261	$-.1130$	2	6313	$-.0086$	1	2
U	8436	$+.0089$	1	3344	$+.1304$	2	8441	$+.0406$	1	3349	$+.1344$	2	2
V	2525	$+.0631$	2	1065	$-.1620$	1	2350	$+.0647$	2	1055	$+.0507$	1	1
W	4181	$-.0180$	1	1792	$-.0191$	1	4054	$-.0207$	1	1792	$-.1347$	2	1
X	1870	$-.0310$	1	742	$-.0080$	1	1832	$-.0438$	1	742	$-.0487$	2	1
Y	6457	$+.0093$	1	2762	$+.0252$	1	6225	$+.0163$	1	2732	$+.0178$	1	1

The problem of stable conditions, or lack of them, further complicates the problem of setting meaningful confidence intervals around projections yielded by regression analysis. Research in this area, perhaps incorporating some of the principles of conditional probability, should be encouraged.

REFERENCES

1. Brown, B.W., Jr.; Savage, R.I., Methodological Studies in Educational Attendance Prediction, Department of Statistics, University of Minnesota, Minneapolis, September 1960.

2. Cochrane, D.; Orcutt, G.H., "Application of Least-squares Regression to Relationships Containing Auto-correlated Error Terms," Journal of the American Statistical Association, 44:32-61, 1949.

3. Johnson, J., Econometric Methods, McGraw-Hill Book Company, Inc., New York, 1960.

4. Larson, K.G.; Strevell, W.H., "How Reliable are Enrollment Forecasts," School Executive, 71:24-28, February 1952.

5. Webster, W.J., The Applicability of Selected Ratio and Least-squares Regression Analysis Techniques to the Prediction of Future Educational Attendance Patterns, PhD dissertation, Michigan State University, East Lansing, 1969.

DIRECTIONS FOR J.E.E. CONTRIBUTORS

The Journal of Experimental Education publishes specialized or technical education studies, treatises about the mathematics or methodology of behavioral research, and monographs of major current research interest.

ABSTRACT REQUIRED

An abstract of not more than 120 words must accompany each manuscript. It should precede the text, be designated *ABSTRACT,* and conform to the following standards:

1. In a research paper, include statements of (a) the problem, (b) the method, (c) the data, and (d) the conclusions.

2. In a review or discussion article, state the topics covered and the central thesis.

3. Standard abbreviations may be used freely if the meaning is clear. Use complete sentences and do not repeat information which is in the title.

TEXT

Usually research articles should have a clearly organized order of presentation. The following elements and their order is a guide and is not intended to be prescriptive.

The Problem. The nature, scope, and significance of the problem should be presented.

Related Research. Presentation and discussion of related research should be selective and should include references essential to clarify the problem under examination.

Methodology. This section should consist of hypotheses, description of the sample and sampling procedures, discussion of the variables, description of the data-gathering instruments (if well-known, their description may be omitted), and general description of the statistical procedures.

Presentation and Analysis of Data. Analysis of the data and conclusions about the hypotheses should be more than mere presentation. Rival hypotheses and significance for related studies and educational theory should be considered. Summary data (means, standard deviations, frequencies, correlation coefficients, etc.) should be presented in tabular or graphic form. Discussion of these data should be interpretive.

Summarizing Statements. A summary of conclusions and implications for education may supplement the abstract.

STYLE

Certain suggestions are offered here to achieve uniformity in publication style and to conserve the time and energy of the author and editorial staff in the preparation copy. *A Manual on Writing Research,* 1962, and *Manual of Form for Theses and Term Reports,* 1962, by Kathleen Dugdale, the Indiana University Bookstore, Bloomington, may be used as style manuals in preparation manuscripts.

Two Copies Required. Copies should be double spaced with wide margins. Typed copies are preferred, but dittoed mimeographed copies will be accepted if they are legible.

Subheads. Articles are frequently improved by the judicious use of subheads. However, avoid use of the title, INTRODUCTION, for a lead section.

Title. Try to use a short title, preferably no more than ten words. Avoid superfluous phrases, such as "A Comparison of . . .," "A Study of . . .," and "The Effectiveness of . . ."

Tables. Prepare tables precisely as they are to appear in *The Journal,* caption each with a brief and to-the-point title, and number consecutively with arabic numerals: *Table 3.* INTERCORRELATION MATRIX, VARIABLES OPTIMALLY ORDERED.

Figures. Draw any graphs and charts in black ink on good quality paper, title each, and number consecutively, thus: *Figure 4.* SCHOOL ENROLLMENT. Send the original copy. We cannot accept mimeographed figures or charts. Data should not be framed, and ordinates and abscissas should be shortened to occupy as little space as possible.

Technical Symbols. All symbols, equations, and formulas must be clearly represented. Differentiate between numbers and letters (zero and the letter o; one and the letter l, etc.) Identify hand-drawn Greek letters in the margin. Align subscripts carefully.

Footnotes. Avoid explanatory footnotes by incorporating their content in the text. However, for essential footnotes, identify them with consecutive superscripts, as book,[2] study,[3] etc., and list the footnotes in a section, entitled FOOTNOTES, at the end of the text, but preceding the REFERENCES.

References. References should be listed alphabetically according to the author's last name at the end of the manuscript. Each entry should be numbered. Citation of a reference in the text should be made with this number, as (2); or if specific pages are to be indicated, as (2:245-7).

Illustrations of article and book references:

1. Bittner, Reign H. and Wilder, Carlton E., "Expectancy Tables: A Method of Interpreting Correlation Coefficients," *The Journal of Experimental Education,* 14:245–52, March 1946.

2. Thomson, Godfrey, H., *The Factorial Analysis of Human Ability,* Houghton Mifflin Co., Boston, 1950, 383 pp.

COSTS

The publisher charges a contributor's fee of $6 per printed page of approximately 1,200 words, billed upon publication. Each contributor will receive 10 complimentary copies of the issue in which his article appears. Reprints are charged at cost, and a price schedule will be sent to each contributor.

PROOFREADING

We will send you proofs for correction (with instructions for handling). Any major changes made in the proofs that were not incorporated in your original copy will be an added expense to you. (Errors that we make, naturally, will be at our expense.)
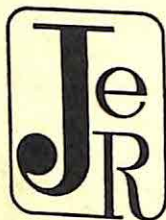
Keep an exact carbon copy of your manuscript so that if a question arises the editors can refer you to specific pages, paragraphs, or lines for clarification by letter.

SEND MANUSCRIPTS TO
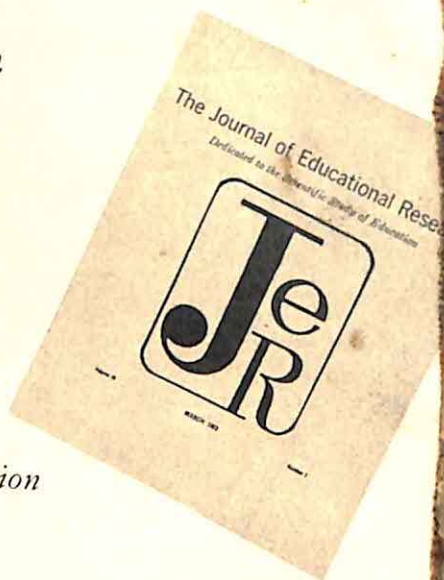
John Schmid, Department of Research and Statistical Methodology, University of Northern Colorado, Greeley, Colorado 80631.